IMMM 2011

The First International Conference on Advances in Information Mining and Management

ISBN: 978-1-61208-162-5

October 23-29, 2011

Barcelona, Spain

IMMM 2011 Editors

Ulrich Norbisrath, University of Tartu, Estonia,

Pascal Lorenz, University of Haute Alsace, France

IMMM 2011

Foreword

The First International Conference on Advances in Information Mining and Management [IMMM 2011], held between October 23 and 29, 2011 in Barcelona, Spain, started 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 2011 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to IMMM 2011. 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 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that IMMM 2011 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 historic charm of Barcelona, Spain.

IMMM 2011 Chairs:

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

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

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

Special Area Chairs on Data Management
Robert Wrembel, Poznan University of Technology, Poland

Special Area Chair on Special Mining
Yulan He, Knowledge Media Institute / The Open University, UK

Special Area Chair on Semantic Data Handling
Stefan Brüggemann, OFFIS - Institute for Information Technology, Germany

Special Area Chair on Databases
Lena Strömbäck, Linköpings Universitet, Sweden

Special Area Chair on Cloud-based Mining
Roland Kübert, High Performance Computing Center Stuttgart / Universität Stuttgart, Germany

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

IMMM 2011

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 2010 Technical Program Committee

Ahmet Aker, University of Sheffield, UK
Zaher Al Aghbari, University of Sharjah, UAE
Andreas S. Andreou, Cyprus University of Technology, Cyprus
César Andrés Sánchez, Universidad Complutense de Madrid, España
Avi Arampatzis, Democritus University of Thrace, Greece
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Philip Azariadis, University of the Aegean - Syros, Greece
Barbara Rita Barricelli, Università degli Studi di Milano, Italy
Alan Barton, NRC, Canada
Shariq Bashir, Vienna University of Technology, Austria
Grigorios N. Beligiannis, University of Ioannina - Agrinio, Greece
Jorge Bernardino, ISEC - Institute Polytechnic of Coimbra, Portugal
Gloria Bordogna, CNR IDPA - Dalmine, Italy
Stefan Brüggemann, OFFIS - Institute for Information Technology, Germany
Wray Buntine, NICTA, Canberra, Australia
Miroslav Bures, Czech Technical University in Prague, Czech Republic
Alain Casali, Aix-Marseille University, France
Sukalpa Chanda, Gjøvik University College, Norway
Jiann-Liang Chen, National Taiwan University of Science and Technology, Taiwan
Yili Chen, Monsanto - St. Louis, USA
Max Chevalier, ITIT, France
Stelvio Cimato, Università degli studi di Milano - Crema (CR), Italy
Tharam Dillon, Curtin University of Technology - Perth, Australia
Qin Ding, East Carolina University - Greenville, USA
Aijuan Dong, Hood College - Frederick, USA
Nikolaos Doulamis, National Technical University of Athens, Greece
Feng Gao, Microsoft, USA
Manuel Gil Pérez, University of Murcia, Spain
Richard Gunstone, Bournemouth University, UK
Allan Hanbury, Information Retrieval Facility - Vienna, Austria
Rania Hatzi, Harokopio University of Athens, Greece
Yulan He, Knowledge Media Institute / The Open University, UK
Andreas Holzinger, Institute for Medical Informatics, Statistics and Documentation (IMI) / Medical University Graz (MUG), Austria
Gilles Hubert, IRIT/University of Toulouse, France
George Ioannidis, IN2 search interfaces development Ltd., UK
Masoumeh Tabae Izadi, Research Institute of McGill University Health Center & School of Computer Science, McGill University, Canada
Ken Kaneiwa, Iwate University, Japan
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Michal Krátký, VŠB - Technical University of Ostrava, Czech Republic
Roland Kübert, High Performance Computing Center Stuttgart / Universität Stuttgart, Germany
Lynda Tamine Lechani, IRIT, France
Kuan-Ching Li, Providence University, Taiwan
Longzhuang Li, Texas A&M University-Corpus Christi, USA
Xuelong Li, Chinese Academy of Sciences, China
Qing Liu, CSIRO, Australia
Shuai Ma, University of Edinburgh, UK
Stephane Maag, TELECOM SudParis, France
Thomas Mandl, University of Hildesheim, Germany

Francesco Marcelloni, University of Pisa, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Arturas Mazeika, Max-Planck-Institut für Informatik, Germany
Johannes Meinecke, SAP AG / SAP Research Center Dresden, Germany
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
Charalampos Moschopoulos, University of Patras, Greece
Josiane Mothe, IRIT-Toulouse, France
Henning Müller, University Hospitals of Geneva, Switzerland
Michael Oakes, University of Sunderland and Uni Computing - Bergen, UK
Kok-Leong Ong, Deakin University, Australia
Yoseba Penya, University of Deusto - DeustoTech (Basque Country), Spain
Nathalie Pernelle, LRI-Paris Sud University, France
Keith Phalp, Bournemouth University, UK
Ioannis Pratikakis, Democritus University of Thrace - Xanthi, Greece
Malte Ressin, Centre for Internationalisation and Usability / Thames Valley University - London, UK
Igor Ruiz-Agundez, University of Deusto / Basque Country, Spain
Antonio Sarasa-Cabezuelo, Complutense University of Madrid, Spain
Hossein Sharif, University of Portsmouth, UK
Simeon Simoff, University of Western Sydney, Australia
Lena Strömbäck, Linköpings Universitet, Sweden
Lynda Tamine-Lechani, Université Paul Sabatier - Toulouse, France
Yi Tang, Chinese Academy of Sciences, China
Xiaohui (Daniel) Tao, Queensland University of Technology, Australia
Olivier Teste, Université de Toulouse / ITIR, France
Ulrich Thiel, IPSI-Fraunhofer, Germany
Konstantinos Tserpes, Harokopio University of Athens, Greece
Chrisa Tsinaraki, Technical University of Crete, Greece
M.N. Vrahatis, University of Patras, Greece
Elizabeth Baoying Wang, Waynesburg University, USA
Qi Wang, University of Science and Technology of China (USTC), China
Wendy Hui Wang, Stevens Institute of Technology - Hoboken, USA
Robert Wrembel, Politechnika Poznanska, Poland
Hao Wu, Yunnan University - Kunming, P. R. China
Pingkun Yan, Philip Research North America, USA
Wei Zhang, Microsoft, USA

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

Mining Epidemiological Data Sources in H1N1 Pandemic Using Probabilistic Graphical Models 1
Masoumeh Tabaeh Izadi, David Buckeridge, and Katia Charland

Comment-guided Learning: Bridging the Knowledge Gap between Expert Assessor and Feature Engineer 7
Xiang Li, Wen-Pin Lin, and Heng Ji

Data Preparation in the MineCor KDD Framework 16
Christian Ernst and Alain Casali

Exploiting Background Information Networks to Enhance Bilingual Event Extraction Through Topic Modeling 23
Hao Li, Heng Ji, Hongbo Deng, and Jiawei Han

Mining Ice Hockey: Continuous Data Flow Analysis 31
Adam Hipp and Lawrence Mazlack

Knowledge Management System IMMPETUS (KMSI) - Connoisseur 37
Osman Ishaque, Mahdi Mahfouf, George Panoutsos, and Lucian Tipi

Optimising Parameters for ASKNet: A Large Scale Semantic Knowledge Network Creation System 42
Brian Harrington and Simon Kempner

Semi-Automated Semantic Annotation for Semantic Advertising Networks 48
Aseel Addawood and Lilac Al-Safadi

Mining Cross-document Relationships from Text 55
Petr Knoth and Zdenek Zdrahal

Analyzing the Use of Word Graphs for Abstractive Text Summarization 61
Elena Lloret and Manuel Palomar

Improving Email Management 67
Tonu Tamme, Ulrich Norbisrath, Georg Singer, and Eero Vainikko

A Redundant Bi-Dimensional Indexing Scheme for Three-Dimensional Trajectories 73
Antonio d'Acierno, Alessia Saggese, and Mario Vento

Reasoning on High Performance Computing Resources 79
Axel Tenschert and Pierre Gilet

Pervasive Ad hoc Location Sharing To Enhance Dynamic Group Tours 85

Markus Duchon, Julian Kopke, Michael Durr, Corina Schindhelm, and Florian Gschwandtner

Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs 91
Jingjing Liu, Alice Li, and Stephanie Seneff

FATS: A Framework for Annotation of Travel Blogs Based on Subjectivity 97
Inmaculada Alvarez de Mon y Rego and Liliana Ibeth Barbosa Santillan

The Migraine Radar - A Medical Study Analyzing Twitter Messages? 103
Dirk Reinel, Sven Rill, Jorg Scheidt, and Florian Wogenstein

Selecting Data Mining Model for Web Advertising in Virtual Communities 107
Jerzy Surma and Mariusz Lapczynski

Large-Scale Association Rule Discovery from Heterogeneous Databases with Missing Values using Genetic Network Programming. 113
Eloy Gonzales, Takafumi Nakanishi, and Koji Zettsu

New Solution for Extracting Inductive Learning Rules and their Post-Analysis 121
Rein Kuusik and Grete Lind

An Equivalence Class Based Clustering Algorithm for Categorical Data 127
Qingbao liu, Wanjun Wang, Su Deng, and Guozhu Dong

Exploiting Student Intervention System Using Data Mining 131
Samia Oussena, Hyensook Kim, and Tony Clark

Supporting Global Design Through Data Mining and Localization 138
Barbara Rita Barricelli and Malte Ressin

ArmSquare: An Association Rule Miner Based on Multidimensional Numbered Information Spaces 143
Iliya Mitov, Krassimira Ivanova, Benoit Depaire, and Koen Vanhoof

Review of Shape-based Similarity Algorithms and Design Retrieval Methods for Computer-aided Design and Manufacturing 149
Leila Zehtaban and Dieter Roller

Mining Information Retrieval Results Significant IR parameters 156
Jonathan Compaore, Adji Mairam Gueye, Sebastien Dejean, Josiane Mothe, and Joelson Randriamparany

Mining Literal Correlation Rules from Itemsets 162
Alain Casali and Christian Ernst

Designing Cost-sensitive Fuzzy Classification Systems Using Rule-weight 168
Mansoor Zolghadri Jahromi and Mohammad Reza Moosavi

Mining Epidemiological Data Sources in H1N1 Pandemic Using Probabilistic Graphical Models

Masoumeh Izadi David Buckeridge Katia Charland
mtabae@cs.mcgill.ca, david.buckeridge@mcgill.ca, katia.charland@mcgill.ca
Clinical and Health Informatics Research Group
McGill University, 1140 Pine Avenue
Montreal, QC, Canada

***Abstract*—It is generally difficult to estimate disease prevalence or true infection probabilities because these are not observable quantities. However, these parameters can be estimated from available data sources that can provide partial indications of the true incidence of infected cases or prevalence rates. However, building a construct capable of incorporating data from these various sources in a coherent manner is not trivial. In addition, the prevalence of an infectious strain must be estimated in a timely manner. For instance, in an epidemic, this estimate must be obtained within a day or so. We propose to use dynamic Bayesian networks from the class of probabilistic graphical models in order to identify probabilistic relationships between different data streams. This is an initial step towards building a framework that can support data integration and real-time estimation of disease prevalence. Our preliminary results on data sources related to H1N1 pandemic show that the proposed models generalize well.**

Key words– data integration; Bayesian networks; time series analysis; surveillance of infectious disease

I. INTRODUCTION

Infectious disease outbreaks result in high human and financial costs. Respiratory and gastrointestinal infectious diseases, in particular, are among the most prevalent types of infections encountered in routine public health practice. The rapid emergence of the novel pandemic (H1N1) 2009 influenza virus in the spring of 2009 was the most recent example with international concern. This pandemic resulted in more than 18,000 deaths since it appeared in April 2009 [1]. Due to the continued threat of influenza and recognizing the importance of methodological advances to estimate the number of infected cases, building models that provide a good level of understanding of the available data is crucial. Several streams of data such as visits to emergency departments, sales of over the counter drugs, calls to health information lines, and admission to hospitals are routinely used for monitoring outbreaks. In addition, with the advances in research on discovering new sources of data for monitoring of infectious diseases, more emerging data streams become available. However, majority of surveillance systems responsible for monitoring these data treat the sources separately or combine them in an ad hoc fashion.

Combining the data sources can increase statistical power of the data and alleviate biases due to confounding and missing values, in general. Building an architecture to fuse data from different sources in a way that can be easily used for reasoning and prediction is not always easy. Moreover, the desired architecture must be scalable, easily updated, and extensible. Classical approaches to time-series prediction includes linear models such as ARIMA (autoregressive integrated moving average), ARMAX (autoregressive moving average exogenous variables model) [2], [3] and Nonlinear models such as neural networks, decision trees. Problems with these approaches include the fact that it is difficult to incorporate prior knowledge and to integrate multi-dimensional sources into these models. We address this problem using probabilistic graphical models which can be used as appropriate tools for data mining.

Probabilistic graphical models are represented by a graph with nodes and links. The main advantage of these models for data mining and analysis is that the graph structure is used to discover a joint probability distribution for any number of known and unknown quantities simultaneously. Bayesian networks (BNs) and hidden Markov models (HMMs) are among the most popular forms of these models. Both models provide promising methodologies for encoding relations among a large number of random variables based on conditional independence property and are easy to represent real-world problems of high degree of complexity. A generalization over these two models is known as dynamic Bayesian networks (DBNs). DBNs generalize Bayesian networks to model temporal relations and generalize HMMs to model interdependencies between observations.

Our objective is to create a DBN as a unified model to mine different data streams for their interrelationships and to use this model for inference and predictions on data sources used in routine biosurveillance. Another important issue we would like to address is the problem of timeliness. This is specially important in the case of epidemics to have estimates of future counts rapidly. In this paper, we show that there is no need to wait for weeks or even a week in order to estimate the counts of important epidemiological data in future. To further elucidate upon the concept of applicability

 ISBN: 978-1-61208-162-5

of DBNs in this context, a case study is persuaded in this research based on available data sources which carry information related to the infected incidence rate of H1N1 over the pandemic period. For the illustration purposes, in this paper we focus on the data from the island of Montreal, Quebec. Through collaboration with the department of public health in Montreal we had access to data sources such as counts of emergency department visits, calls to health-information lines, vaccination, and hospitalization. There are known qualitative relationships between infection rate or Influenza Like Illnesses (ILI) incidences and a variety of other data sources. For instance, vaccination would reduce the rate of infected cases. Several quantitative relationship between some of these data are also known as domain knowledge. For instance, flu infection makes almost one third of the ILI visits. While very useful, these distributed pieces of information alone are not sufficient to establish a comprehensive model. DBNs are capable of incorporating such domain knowledge in their structure while they build on the knowledge discovered by the data. The steps in the reasoning and prediction by these models will be illustrated through the H1N1 case study in this paper.

II. Dynamic Bayesian Networks

A Bayesian network is a special type of probabilistic graphical models that is represented by a Directed Acyclic Graph (DAG). The DAG explicitly represents independence relationships among random variables. A DAG contains nodes for each random variable and a link between any two statistically correlated nodes. The node originating the directed link is a parent and the terminating node a child. Each node contains a conditional probability table (CPT) that describes the relationship between the node and its parents. If the topology is unknown, i.e., the independence relations among the random variables is unknown, an appropriate structure must be elicited from the data. Automatically learning the structure of a Bayesian network DAG from data is a well-researched but computationally difficult problem [4], [5], [6], [7]. A function is used to score a network with respect to the training data, and a search method is used to look for the network with best score. Different scoring metrics and search methods have been proposed in the literature. The scoring functions used to select models are based on the likelihood function of a model given the data or the logarithm of this function. Since the associated search space is exponentially large, local search-based approaches, which iteratively consider local changes (adding, deleting, and reversing an edge) to the network structure, are usually used to find the best network. This type of search is very useful when dealing with large data sets because of its computational efficiency. One of the most popular search strategies due to its simplicity and good performance [8], [9], [10] in this context is greedy hill-climbing search which starts from an empty graph and gradually improve it by applying the highest scoring single edge addition or removal available. Once the DAG is learned, the parameters of the model (CPTs) need to be specified or directly learned from data. CPTs identify the probabilities of the child being in any specific values given the values of its parents. Parameter learning in Bayesian networks mainly considers maximum likelihood estimation of the model given the data and it is performed through an expectation maximization process. See [7], [11], [6], [12] for parameter learning methods in Bayesian networks. The advantage of DBNs is being able to represent uncertainties, dependencies and dynamics exhibited in different time series. A DBN consists of a finite number of BNs called slices, where each slice corresponds to a particular time instant. BNs corresponding to successive instants are connected through arcs that represent how the state of a random variable changes over time. A DBN is generally assumed to satisfy the Markov property. It is generally assume that the dependencies between the slices of a DBN and their strength do not change over time. Therefore, a DBN can be described by at most a k-slice network (for a k-order Markov domian). DBNs have been applied in a variety of applications from activity recognition and monitoring to medical diagnosis and fault or defect detection. This is the first time that this framework is used for mining in epidemiological data. Ideally, we should be able to learn and discover the probabilistic relationships between data streams through structure learning in DBNs. However, when the system consist of many data streams and in particular when it is partially observed, structure learning in DBNs becomes computationally intensive. This is due to the fact that the space of possible models is so huge that it will be necessary to use strong prior domain knowledge to make the task tractable.

III. Discovering probabilistic relationships between data streams

One practical approach to discover the relationship between time series is to use statistical techniques to learn about the temporal relationships such as lag-lead relationships among the data sources first, combined that with domain knowledge, and then use this information to construct the required DBN. A popular technique in statistics is used for discovering the relationships between time series data or more generally on sequential data, namely: Wavelet Coherency Analysis (WCA) [13], [14].

Wavelet analysis is a useful mathematical technique for analyzing time-series data and periodicities. The wavelet analysis has found many applications in studying longitudinal data [15], [16], [17]. The wavelet coherence is especially useful in highlighting the time and frequency intervals where two time-series have a strong interaction. Such a spectral analysis should be done in an exhaustive way to find the best fit.

 ISBN: 978-1-61208-162-5

The Coherence is defined as the cross-spectrum normalized to an individual power spectrum. It is a number between 0 and 1, and gives a measurement of the cross-correlation between two time-series and a frequency function. The wavelet squared coherency is a measure of the intensity of the covariance of the two series in time-frequency space [18]. It is used to identify frequency bands within which two time series are co-varying. The WCA can provides insight into the temporal relationships to explore in the Bayesian network setting. This is done via the computation of time-frequency maps of the time-variant coherence [15].

IV. Experimental Evaluation

We used and evaluated DBNs in the context of data integration from different sources which partially indicate the pattern of Influenza H1N1 infection. Although, conventionally DBNs are based on first-order Markov processes (i.e. they can be implemented by one-step temporal relationships between two static BNs for only two consecutive time slices), we observed that the data sources we have in hand may potentially indicate more than one step lag between the time series. Therefore, embedding of this particular information into a DBN formulation requires a k-order Markov process for representing a k-layer network, where k indicates the maximum lag between the time-series.

The experiments reported here are based on the data presented in the next section. We learned DBN models from the data in a variety of settings, and compared them with respect to their performance in predicting observable data streams. The main purpose of this phase of the research is to understand how well DBNs can represent the whole processes, how many observations are required, and what sorts of observations are most useful. In all our experiments, we enforced the presence of the arcs in the DBN network structure based on the suggested settings by WCA, or BN structure learning. In performing the BN structure learning, we followed a similar strategy to what suggested by [19]. For each data source, we selected the variables observed at $t, t+1..., t+10$ days and performed hill climbing search to find the network with the best score.

A. Data

Through collaborations with the department of public health in Montreal, we had access to five different data sources. These data sources include: daily counts of emergency department visits (ED), daily counts of calls to health information lines,Info-Sante, (IS), weekly counts of H1N1 vaccination, weekly counts of confirmed cases of H1N1 through lab tests, and weekly counts of admission to the hospitals. Since the data sources have different resolutions in time and have different significance in predicting the number of infected cases, we are only considering the daily time-series of ED and IS, preliminary. Emergency department visits may well estimate incidence of influenza.

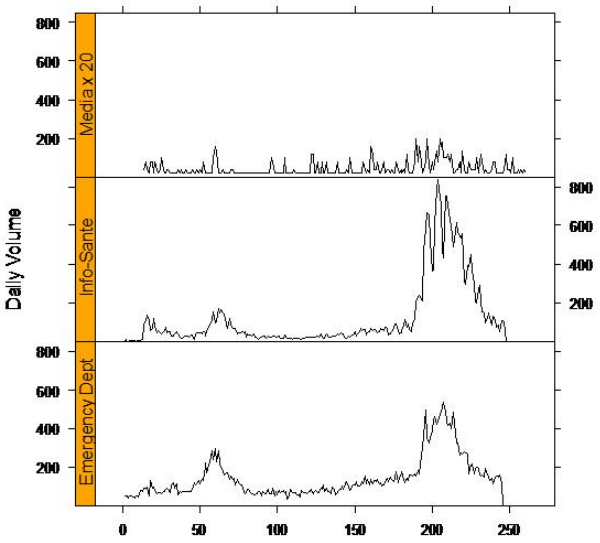


Figure 1. Three data sources of media reports, calls to InfoSante, and emergency department visits from top to bottom.

We can combine emergency department triage data with the telephone survey to characterize the effectiveness of incidence estimation. We aggregated visits for ILI by age group, sex, and day of visit. Similar to the ED data, the IS data can be used to estimate influenza incidence. We aggregated ILI calls by age group, sex, and day of call.

Media reports of deaths from pH1N1 were considered important because of their pronounced effect on the utilization of health services, thus media reports were filtered for content. We also extracted the Media data from the Healthmap [20] on a daily basis.Figure 1 shows the total daily counts of H1N1 media reports about Montreal during the period of April 28, 2009 to December 16, 2009 in the top graph. The second graph illustrate the total daily calls to Info-Sante, and the third graph shows the total daily counts of emergency department visits during the same period. The arrow points to the time when a 13 year old boy (hockey player) in Ontario died on October 26. There were reports of his funeral at around November 4 (t=203 on the time axis). This precedes, by 1 day, the sharp spike in Info-sante calls.

B. Results

The extent of the temporal relationship between IS and ED series data was estimated using WCA in Figure 2. Our results in Figure 2 shows about 2-4 day lead or lag. There is a phase change at around Nov 2, in the second wave. We are able to see a predictable relationship during seasonal influenza (with IS leading ED by approximately 4 days), but during the pandemic (and especially the second wave)

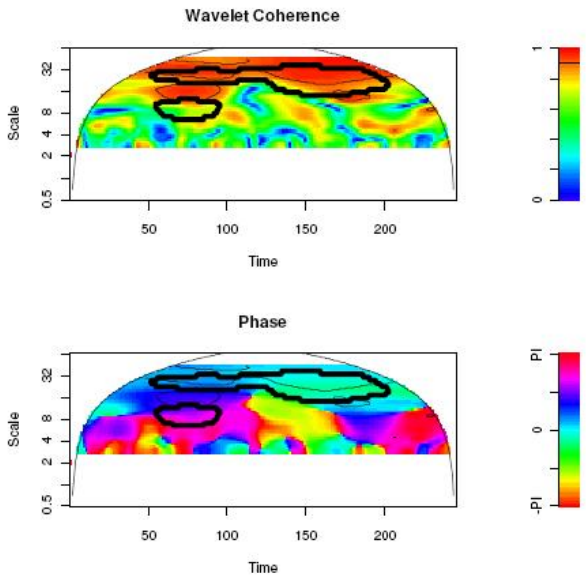


Figure 2. Wavelet coherency analysis for two data sources ED and InfoSante.

the relationship was less predictable. We speculated that it is possibly due to media influence.

In this research, we aimed to learn DBN models that generalize well. The generalization ability of a model G is interpreted as the expected predictive accuracy for the next time series, D_{T+1}. We evaluated the DBN model for prediction accuracy of important observations in time series IS and ED through cross validation techniques. The first set of experiments involved learning DBNs of different complexities. Once trained, we can use the model to do real-time prediction through approximate inference in BNs.

We used a BN structure learning search over the space of all possible graphs to find the best graph, and we discovered two day lag for ED during the seasonal and pandemic flu 2009. However, for an extended period of time (May 1, 2008 to December 30, 2009), which includes non-pandemic, seasonal, and pandemic flu, we found different dependency relations between the two series by BNs structure learning. As the WCA suggested candidate models with 4 days lag, we also tried to train a DBN model with no-phase difference between IS and ED in a DBN (Figure 3). The structure learning method also found that media reports data can lead the Info-Sante data by one day. However, this relationship only exist during the pandemic period in our data sets (April-December 2009). We also presented the Bayesian network models to the experts in public health surveillance and asked them to assess the face validity of the dependence between the time series. The expert feedback was more in favor of IS leading ED.

We experimented with four DBNs that correspond to the settings suggested by BN structure learning and WCA:

- ED leads InfoSante by 2 days
- No phase difference between ED and InfoSante
- InfoSante leads ED by 2 days
- Media leads InfoSante by one day and Infosante leads ED by 2days

Figure 3 shows the unrolled DBNs for seven time steps (weekly). We can treat the unrolled version of a network as a static BN and apply inference algorithms in BNs. We used cross validation for evaluation of all models. Four fifth of the data was used for training and One fifth of the data was used for testing. In each model we provided the information for today's count on ED and IS and predicted the first to 6th next day's counts on both ED and IS. The second model works actually the best when it is trained and tested on the pandemic period (no more than 11% error in predicting ED).

It should be noted that for all models we considered categorization for all variables. This includes Media $\in \{0, 1-3, 3-7, >7\}$, ED $\in \{0-100, 100-200, 200-300, 300-400, 400-500, >500\}$, and IS $\in \{0-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, >700\}$. The results may vary with changing the categorization.

Although, there exist dependencies between the media data and the IS data, we did not see a significant changes in the prediction results for IS. This can be potentially related to other factors which have not been considered in our model or solely related to the experimental setup we selected for these evaluations including the discretization levels of the Media and IS variables and the information provided for reasoning at each time.

V. Conclusions and Future Work

Monitoring epidemiological data is critical for detecting epidemics and for guiding control measures. During the H1N1 pandemic, the Direction de sante publique de Montreal collected data from multiple sources to describe H1N1 influenza infection and associated health care utilization. None of these data sources alone are believed to measure the incidence of H1N1 influenza accurately. In this paper, we proposed a probabilistic graphical model to different heterogeneous data and discover meaningful information these data exhibit. We showed how a DBN model can be used for generating short-term predictions of real-time surveillance data. The estimates are also timely. We showed that only the order of one to two days required in order to estimate future counts in the studied data sources These estimates will be eventually useful in forecasting the spread of H1N1 influenza.

We will continue our investigations for choosing a better DBN structure. We plan to evaluate all lags (plus/minus) 4 and pick the one with the best prediction power. We did not consider the complete DBN model to predict the number of infected cases of H1N1 in this paper. After reaching a good DBN model for integration of data sources, we plan to

 ISBN: 978-1-61208-162-5

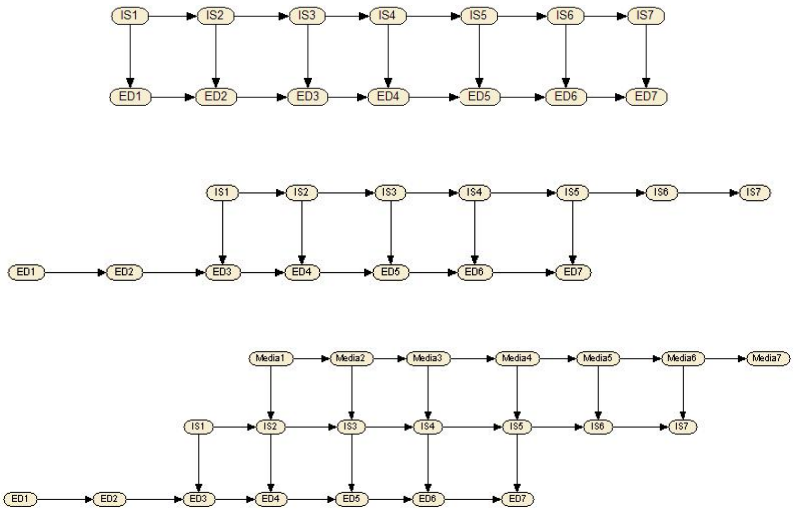


Figure 3. (a) No phase difference between ED and InfoSante, (b) Infosante leads ED by 2days, and (c) Media leads InfoSante by one day and Infosante leads ED by 2days

Table I
COMPARISON OF THE PERFORMANCE OF DIFFERENT DBN MODELS IN PREDICTING INFORMATION-SANTE DATA IN THE NEXT SIX FOLLOWING DAYS.

Model	**error%**					
	Day1	Day2	Day3	Day4	Day5	Day6
ED leads 1-day	19.49	21.19	22.03	26.72	29.03	29.9
IS leads 2-days	18.68	23.37	25.64	26.22	28.81	29.06
Zero-phase difference	18.68	23.37	25.64	26.22	28.81	29.06
Media-effect	18.24	24.21	26.72	27.65	29.31	32.59

Table II
COMPARISON OF THE PERFORMANCE OF DIFFERENT DBN MODELS IN PREDICTING ED DATA IN THE NEXT SIX FOLLOWING DAYS.

Model	**error%**					
	Day1	Day2	Day3	Day4	Day5	Day6
ED leads 1-day	8.47	11.86	14.53	16.1	21.37	24.14
IS leads 2-days	8.47	11.02	12.52	13.33	13.56	18.49
Zero-phase difference	9.32	11.68	12.71	13.64	13.68	16.38

extend the DBN model of observable data sources presented here to what is called an autoregressive hidden Markov models (AHMM) to contain the unobservable infected counts. We can then apply learning algorithms such as Viterbi and Baum-Welch on this hierarchical dynamic Bayesian network just as we can on HMMs to estimate the prevalence of H1N1.

ACKNOWLEDGMENT

The authors acknowledge the contribution of the members of the department of public health in Montreal who agreed to provide data for this research and participated in regular expert meetings. In particular, we would like to thank Lucie Bedard and Robert Allard for invaluable insights for the analysis of our results and in preparation of this paper.

REFERENCES

[1] CIDRAP, "WHO says H1N1 pandemic is over."

[2] H. Burkom, S. Murphy, J. Coberly, and K. Hurt-Mullen, "Public health monitoring tools for multiple data streams," *MMWR*, no. 54, pp. 55–62, August 2005.

[3] B. Reis and K. Mandl, "Time series modeling for syndromic surveillance." *BMC Medical Information Decision Making*, vol. 3, no. 2, 2003.

[4] D. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of bayesian networks is np-hard," *JMLR*, vol. 5, pp. 1287–1330, 2004.

[5] D. Chickering and C. Meek, "Finding optimal bayesian networks," Microsoft Research, Tech. Rep., 2002.

 ISBN: 978-1-61208-162-5

[6] F. A. Jensen, *An Introduction to Bayesian Networks*. Springer, 1996.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[8] I. Tsamardinos, L. Brown, and C. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.

[9] D. Heckerman, D. Geiger, and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, 1995.

[10] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," *Uncertainty in Artificial Intelligence*, pp. 139–147, 1998.

[11] D. Grossman and P. Domingos, "Learning bayesian network classifiers by maximizing conditional likelihood," in *ICML*, 2004.

[12] R. Neapolitan, *Learning Bayesian Networks*. Prentice Hall, 2003.

[13] J. Morlet, G. Arens, I. Foourgeau, and D. Giard, "Wave propagation and sampling theory," *Geophysics*, vol. 47, pp. 203–236, 1982.

[14] S. Mallat, *A Wavelet Tour of Signal Processing*. New York Academic, 1999.

[15] K. Keissar, R. Davrath, and S. Akselrod, "Time and frequency wavelet transform coherence of cardio-respiratory signals during exercise," *Computers in Cardiology*, pp. 733–736, 2006.

[16] T. Li and W. Klemm, "Detection of cognitive binding during ambiguous figure tasks by wavelet coherence analysis of eeg signals," *Pattern Recognition*, pp. 3098–3101, 2000.

[17] A. Grinsted, J. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear Processes in Geophysics*, vol. 11, pp. 561–566, 2004.

[18] C. Torrence and G. Compo, "A practical guide to wavelet analysis," *Program in Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado*, 1998.

[19] P. Sebastiani, K. Mandl, P. Szolovits, I. Kohane, and M. Romain, "Bayesian dynamic model for influenza surveillance," *Journal of the American Statistical Association*, 2006.

[20] J. Brownstein, C. Freifeld, B. Reis, and K. Mandl, "Surveillance sans frontires: Internet-based emerging infectious disease intelligence and the healthmap project," *PLoS Med*, vol. 5, no. 7, p. 151, 2008.

 ISBN: 978-1-61208-162-5

Comment-guided Learning: Bridging the Knowledge Gap between Expert Assessor and Feature Engineer

Xiang Li, Wen-Pin Lin, Heng Ji
Computer Science Department, Queens College and Graduate Center
City University of New York
New York, USA
{jackieiuu729,danniellin,hengjicuny}@gmail.com

***Abstract*—As more and more natural language processing systems utilize human assessment on system responses, it becomes beneficial to discover some hidden privileged knowledge (such as comments) from assessors. We present a simple, low-cost but effective comment-guided learning approach to exploit such knowledge declaratively in an automatic assessor. Our approach only requires a small set of training data, together with comments which are naturally available from human assessment. To demonstrate the power and generality of this approach, we apply the method in two very different applications: name translation and residence slot filling. Our approach achieved significant absolute improvement (15% for name translation and 8% for slot filling) over state-of-the-art systems. It also outperformed previous methods such as Recognizing Textual Entailment (RTE) based fact validation. Furthermore, it can be used as feedback to significantly speed up human assessment.**

Keywords-comment-guided learning; assessment; feature engineering.

I. INTRODUCTION

As an inter-disciplinary area, statistical Natural Language Processing (NLP) requires two crucial aspects: (1) good choice of machine learning algorithms; (2) good feature engineering. For many NLP tasks, feature engineering significantly affects the performance of systems. However, feature engineering is very challenging because it encompasses feature design, feature selection, feature induction and studies of feature impact, all of which are very time-consuming, especially when there are a lot of data or errors to analyze. As a result, in a typical feature engineering process, the system developer is only able to select a representative data set as the development set and analyze partial errors.

On the other hand, recently many NLP tasks have moved from processing hundreds of documents to large-scale or even web-scale data. Once the collection grows beyond a certain size, it is not feasible to prepare a comprehensive answer key in advance. Because of the difficulty in finding information from a large corpus, any manually-prepared key is likely to be quite incomplete. Instead, we can pool the responses from various systems and have human assessors manually review and judge the responses. Assessing pooled system responses as opposed to identifying correct answers from scratch has provided a promising way to generate training data for NLP systems.

However, almost all of the previous NLP systems only utilized the direct assessment results (correct, incorrect, etc.) for training, while ignoring the valuable knowledge hidden in the human assessment procedure. If we consider an NLP system as a "student" while the human assessor as a "teacher", then the "homework grades" (i.e., assessment results) are just a small part of the teacher's role. Besides grading, a teacher also provides explanations about why an answer is wrong, comments about what kind of further knowledge the student can benefit from, and so on. Similarly, when a human assessor makes a judgement, he/she must know the reasons to verify it. As a result, it will cost little extra time for an assessor to write down their comments, because the comments can be naturally derived from a small yet representative sample data set that the assessor has judged. Such assessment results and comments are not available in the test phase. However, since a system tends to make similar types of errors on various data sets, it can always benefit from such comments for new runs.

In this paper, we propose a new and general Comment-Guided Learning (CGL) framework in order to fill in the gap between the expert annotator and the feature engineer (Section 3). This framework aims to encode features with the guidance of comments from human assessors in a re-scoring step. In order to verify the efficacy of this approach, we shall conduct case studies on two distinct application domains: a relatively simple name translation task (Section 4) and a more challenging residence slot filling task (Section 5). Empirical studies demonstrate that with about little longer annotation time, we can significantly improve the performance for both tasks.

II. RELATED WORK

Vapnik [1] proposed to incorporate more of "teacher's role" (i.e., privileged knowledge) into traditional machine learning paradigm, and pointed out that such privileged knowledge may not be available during the test phase. We follow this basic idea and incorporate additional feedback from the comments into assessment, so that we can still

 ISBN: 978-1-61208-162-5

utilize the teacher's role as the final step of the test phase. Vapnik's work aimed to improve a classifier using privileged knowledge by exploiting information from a different classification space to tune the classifier to make better predictions. Their updated classifier is still applied to the same space as the original one. In contrast, we encode the comments from assessors as new features for an automatic assessor. The updated system consisted of the baseline and the assessor is applied to a new classification space (i.e., new test instances).

Castro et al. [2] investigated a series of human active learning experiments. Our experiment of using CGL to speed up human assessment exploited assistance from multiple systems.

Our idea of learning from error corrections is also similar to Transformation-based Error-Driven Learning, which has been successfully applied in many NLP tasks such as part-of-speech tagging [3] and semantic role labeling [4]. In these applications the transformation rules are automatically learned based on sentence contexts at each iteration. However, our applications require global knowledge which may be derived from diverse linguistic levels and vary from one system to the other, and thus it's not straightforward to design and encode transformation templates. Therefore in this paper we choose a more modest way of exploiting the comments encoded by human assessors.

Most of the previous name translation work combined supervised phonetic similarity based name transliteration approaches with Language Model based re-scoring (e.g., [5]). But none of these approaches exploited the feedback from human assessors. There are many other alternative automatic assessment approaches for slot filling. Besides the RTE-KBP validation [6] discussed in the paper, some slot filling systems also conducted filtering and cross-slot reasoning (e.g., [7]; [8]) to improve results.

III. COMMENT-GUIDED LEARNING

A. General Framework

In Table I, we aim at formalizing the mapping of some essential elements in human learning and machine learning for NLP. We can see that among these elements, little study has been conducted on incorporating the comments made by human assessors. In most cases it was not the obligation for the human assessors to write down their comments during assessment. In contrast, the human learning scenario involves more interactions. However, we can assume that any assessor is able to verify and comment on his/her judgement. Based on this intuition we propose a new comment-guided learning (CGL) paradigm as shown in Figure 1. This algorithm aims to extensively incorporate all comments from an old development data set (i.e., "old homework" in human learning) into an automatic correction component. This assessor can be applied to improve the results for a new test data set (i.e., "new homework" in human learning).

Table I
SOME ELEMENTS IN HUMAN LEARNING AND MACHINE LEARNING FOR NLP

Human Learning	Machine Learning for NLP	Examples of existing NLP approaches
student	system	baseline system
teaching assistant	human annotator	
teacher	human assessor	
lectures	training data	
graded homework	assessed system output	
graded homework with comments	assessed system output with comments	None
erroneous homework set	negative samples/errors	transformation based learning
homework review against lecture	system output with background documents	recognizing textual entailment
group study	pooled system responses	voting, learning-to-rank

B. Detailed Implementation

The detailed CGL algorithm can be summarized as follows.

1. The pipeline starts from running the baseline system to generate results. In this step we can also add the outputs from other systems (i.e., classmates in human learning) or even human annotators (i.e., Teaching Assistant (TA) in human learning). We will present one case study on slot filling, which incorporates these two additional elements, and the other case study on name translation which only utilizes results from the baseline system.

2. We obtain comments from human assessors on a small development set D^i. Each time we ask a human assessor to pick up N (N=3 in this paper, the value of 3 was arbitrarily chosen; various in this number of clusters produce only small changes in performance.) random results to generate one new comment. One could impose some predefined format or template restrictions for the comments, such as marking the indicative words as rich annotations and encoding them as features. However, we found that most of the expert comments are rather implicit and even requires global knowledge. Nonetheless, these comments represent general solutions to reduce the common errors from the baseline system.

 ISBN: 978-1-61208-162-5

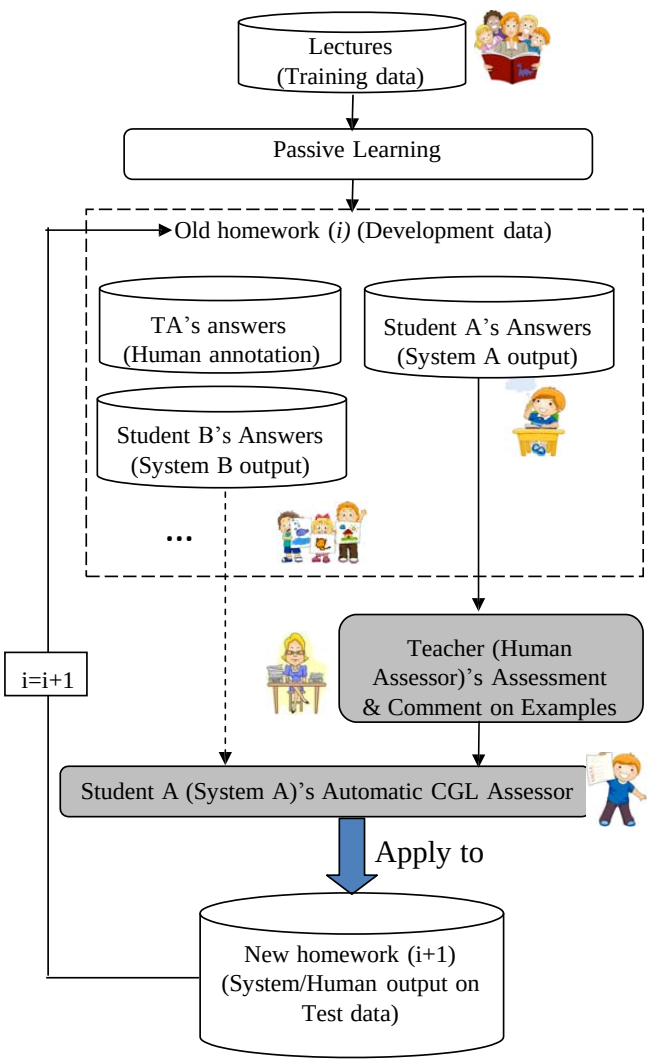


Figure 1. Training a CGL Assessor

3. We encode these comments into features. We then further train a Maximum Entropy (MaxEnt) based automatic assessor A^i using these features. For each response generated from the baseline system, A^i can classify it as correct or incorrect. We choose a statistical model instead of rules because heuristic rules may overfit a small sample set and highly dependent on the order. In contrast, MaxEnt model has the power of incorporating all comments into a uniform model by assigning weights automatically. In this way we can integrate assessment results tightly with comments during MaxEnt model training.

4. Finally, A^i is applied as a post-processing step to any new data set D^{i+1}, and filter out those results judged as incorrect.

The algorithm can be conducted in an iterative fashion. For example, human assessors can continue to judge and provide comments for D^{i+1} and we can update the automatic assessor to A^{i+1} and apply it to a new data set D^{i+2}, and so on.

IV. Name Translation Mining

This section presents the first case study of applying CGL for name translation validation.

A. Task Definition and Baseline System

Name translation is important well beyond the relative frequency of names in a text: a correctly translated passage, but with the wrong name, may lose most of its value. Some recent work explored unsupervised or weakly-supervised name translation mining from large-scale data. For example, Bouma et al. [9] aligned attributes in Wikipedia Infoboxes based on cross-page links; Lin et al. [10] described a parenthesis translation mining method; You et al. [11] applied graph alignment algorithm to obtain name translation pairs based on co-occurrence statistics. However, these approaches suffer from low accuracy and thus it is important to develop automatic methods to evaluate whether the mined name pairs are correct or not. In this paper we focus on validating person name translations by encoding the comments which human assessors made on a small data set. We applied an unsupervised learning approach as described in [12] as our baseline system, to mine name translation pairs from English and Chinese Wikipedia Infoboxes.

B. Comments and Feature Encoding

The detailed comments used for validating name translations are as follows.

- Comment 1: "these two names do not co-occur often" This comment indicates that we can exploit global statistics to filter out some obvious errors, such as "Ethel Portnoy" and "Chen Yao Zu". Using Yahoo! search engine, we compute the co-occurrence, conditional probability and mutual information of a Chinese Name CHName and an English name ENName appearing in the same document from web-scale data with setting a threshold for each criteria.
- Comment 2: "these two names have very different pronunciations"
  Many foreign names are transliterated from their origin pronunciations. As a result, person name pairs (e.g., "Lomana LuaLua" and "Luo Ma Na . Lu A Lu A") usually share similar pronunciations. In order to address this comment, we define an additional feature based on the Damerau-Levenshtein edit distance ([13]) between the Pinyin form of CHName and ENName. Using this feature we can filter out many incorrect pairs, such as "Maurice Dupras" and "Zhuo Ya . Ke Si Mo Jie Mi Yang Si Qia Ya".
- Comment 3: "these two names have different profiles" When human assessors evaluate the name translation pairs, they often exploit their world knowledge. For example, they can quickly judge "Comerford Walker" is not

a correct translation for "Sen Gang Er Lang (Jiro Oka Mori)" because they have different nationalities (one is U.S. while the other is Japan). To address this comment, we define the profile of a name as a list of attributes. Besides using all of the Wikipedia Infobox values, we also run a bi-lingual information extraction (IE) system [14] on large comparable corpora (English and Chinese Gigaword corpora) to gather more attributes for ENName and CHName. For example, since "Nick Grinde" is a "film director" while "Yi Wan . Si Te Lan Si Ji" is a "physicist" in these large contexts, we can filter out this incorrect name pair.

The detailed features converted from the above comments are summarized in Table II.

Table II
VALIDATION FEATURES FOR NAME TRANSLATION

Comments	Features
1	co-occurrence, conditional probability and mutual information of *CHName* and *ENName* appearing in the same document from web-scale data
	conditional probability of *CHName* and *ENName* appearing in the same document from web-scale data
	mutual information of *CHName* and *ENName* appearing in the same document from web-scale data
2	Damerau-Levenshtein edit distance between the Pinyin form of *CHName* and *ENName*
3	overlap rate between the attributes of *CHName* and the attributes of *ENName* according to Wikipedia Infoboxes
	overlap rate between the attributes of *CHName* and the attributes of *ENName* according to IE results of large comparable corpora

C. Data and Scoring Metric

We used English and Chinese Wikipedias as of November 2010, including 10,355,225 English pages and 772,826 Chinese pages, and mined 5368 name pairs. A small set of 100 pairs is taken out as the development set for the human assessor to encode comments. The CGL assessor is then trained and tested on the remaining pairs by 5-folder cross-validation.

It is time consuming to evaluate the mined name pairs because sometimes the human assessor needs Web access to check the contexts of the pairs, especially when the translations are based on meanings instead of pronunciations. We implemented a baseline of mining name pairs from cross-lingual titles in Wikipedia as an incomplete answer key, and so we only need to ask two human assessors (not system developers) to do manual evaluation on our system generated pairs which are not in this answer key (1672 in total). A name pair is judged as correct if both of them are correctly extracted and one is the correct translation of the other. Such a semi-automatic method can speed up evaluation. On average each human assessor spent about 3 hours on evaluation.

D. Overall Performance

Table III shows Precision (P), Recall (R) and F-measure (F) scores before and after applying the CRL assessor on name translation pairs. As we can see from Table III, CGL achieved 28.7% absolute improvement on precision with a small loss (4.9%) in recall. In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on F-measures. The results show that we can reject the hypothesis that the improvements using CGL were random at a 99.8% confidence level.

Table III
OVERALL CGL PERFORMANCE ON NAME TRANSLATION

Apply CGL	P (%)	R (%)	F (%)
before	69.3	1	81.9
after	98.0	95.1	**96.5**

V. SLOT FILLING

In this section, we shall apply CGL to a more challenging task of slot filling and investigate the detailed aspects of CGL by comparing it with other alternative methods.

A. Task Definition

In the slot filling task [15], attributes (or "slots") derived from Wikipedia infoboxes are used to create the initial (or reference) knowledge base (KB). A large collection of source news and web documents is then provided to the slot filling systems to expand the KB automatically.

The goal of slot filling is to collect information regarding certain attributes of a query from the corpus. The system must determine from this large corpus the values of specified attributes of the entity. Along with each slot answer, the system must provide the ID of a document which supports the correctness of this answer.

We choose three residence slots for person entities ("countries_of_residence", "stateorprovinces_of_residence" and "cities_of_residence") for our case study because they are one group of the most challenging slot types for

which almost all systems perform poorly (less than 20% F-measure).

B. Baseline Systems

We use our slot filling system [7] which achieved highly competitive results (ranked at top 3 among 31 submissions from 15 teams) at the KBP2010 evaluation as our baseline. This system includes multiple pipelines in two categories: bottom-up IE based approaches (pattern matching and supervised classification) and a top-down Question Answering (QA) based approach that search for answers constructed from target entities and slot types. The overall system begins with an initial query processing stage where query expansion techniques are used to improve recall. The best answer candidate sets are generated from each of the individual pipelines and are combined in a statistical reranker. The resulting answer set, along with confidence values are then processed by a cross-slot reasoning step based on Markov Logic Networks [16], resulting in the final system outputs. In addition, the system also exploited external knowledge bases such as Freebase [17] and Wikipedia text mining for answer validation.

In order to check how robust the CGL assessor is, we also run it on some other anonymous systems in KBP2010 with representative performance (high, medium and low).

C. Comments and Feature Encoding

The detailed comments used for our slot filling experiment are as follows.

- Comment 1: "this answer is not a geo-political name"
  This comment is intended to address some obvious errors which could not be Geo-political (GPE) names in any contexts. In order to address this comment, we apply a very large gazetteer of GPE hierarchy (countries, states and cities) from the geonames website [2] for answer validation.
- Comment 2: "this answer is not supported by this document"
  Some answers obtained from Freebase may be incorrect because they are not supported by the source document. Answer validation was mostly conducted on the document basis, but for the residence slots we need to use sentence-level validation. In addition, some sentence segmentation errors occur in web documents. To address this comment, we apply a coreference resolution system [12] to the source document, and check whether any mention of the query entity and any mention of the candidate answer entity appear in the same sentence.
- Comment 3: "this answer is not a geo-political name in this sentence"
  Some ambiguous answers are not GPE names in certain contexts, such as "European Union". To address this comment, we extract the context sentences including the query and answer mentions, and run a name tagger [18] to verify the candidate answer is a GPE name.
- Comment 4: "this answer conflicts with this system/other system's output"
  When an answer from our system is not consistent with another answer which appears often in the pooled system responses, this comment suggests us to remove our answer. In order to address this comment, we implemented a feature based on hierarchical spatial reasoning. We conduct majority voting on all the available system responses, and collect the answers with global confidence values (voting weights) into a separate answer set ha. Then for any candidate answer a, we check the consistency between a and any member of ha by name coreference resolution and part-whole relation detection based on the gazetteer of GPE hierarchy as described in Comment 1. For example, if "U.S." appears often in ha we can infer "Paris" is unlikely to be a correct answer for the same query; on the other hand if "New York" appears often in ha we can confirm "U.S." as a correct answer.

The detailed features converted from the above comments are summarized in Table IV.

Table IV
VALIDATION FEATURES FOR SLOT FILLING

Comments	Features
1	whether the answer is in the geo-political gazetteer
2	whether any mention of the query entity and any mention of the answer entity appear in the same sentence using coreference resolution
3	whether the answer is a GPE name by running name tagging on the context sentence
4	whether the answer conflicts with the other answers which received high votes across systems using inferences through the GPE hierarchy

D. Data and Scoring Metric

During KBP2010, an initial answer key annotation was created by LDC through a manual search of the corpus, and then an independent adjudication pass was applied by LDC human assessors to assess these annotations together with pooled system responses to form the final gold-standard answer key. We incorporated the assessment comments for

 ISBN: 978-1-61208-162-5

our system output on a separate development set (182 unique non-NIL answers in total) from KBP2010 training data set to train the automatic assessor. Then we conduct blind test on the KBP2010 evaluation data set which includes 1.7 million newswire and web documents. The final answer key for the blind test set includes 81 unique non-NIL answers for 49 queries.

The number of features we can exploit is limited by the unknown restrictions of individual systems. For example, some other systems used distant learning based answer validation and so could not provide specific context sentences. Since comment 2 and comment 3 require context sentences, we trained one assessor using all features and tested it on our own system. Then we trained another assessor using only comment 1 and 4 and tested it on three other systems representing different levels of performance.

Equivalent answers (such as "the United States" and "USA") are grouped into equivalence classes. Each system answer is rated as correct, wrong, or redundant (an answer which is equivalent to another answer for the same slot or an entry already in the knowledge base). Given these judgments, we calculate the precision, recall and F-measure of each system, as defined in [15].

E. Overall Performance

Table V shows the slot filling scores before and after applying the CGL assessors (because of the KBP Track requirements and policies, we could not mention the specific names of other systems). The Wilcoxon Matched-Pairs Signed-Ranks Test show we can reject the hypothesis that the improvements using CGL over our system were random at a 99.8% confidence level. It also indicates that the features encoded from comment 2 and comment 3, which require intermediate results such as context sentences helped boost the performance about 3.4%. We can see that although the other high-performing system may have used very different algorithms and resources from ours, our assessor still provided significant gains. Our approach improved the precision on each system (more than 200% relative gains) with some loss in recall. Since most comments focused on improving precision, F-measure gains for moderate-performing and low-performing systems were limited by their recall scores. This is similar to the human learning scenario where students from the same grade can learn more from each other than from different grades. In addition, the errors removed by our approach were distributed equally in newswire (48.9%) and web data (51.1%), which indicates the comments from human assessors reached a good degree of generalization across genres.

F. Cost and Contribution of Each Comment

The comments from the CGL assessor may reflect different aspects of the system. Therefore it will be interesting to investigate what types of comments are most useful and not costly. We did another experiment by applying one comment at a time into the assessor. Table VI shows the results along with the cost of generating and encoding each comment (i.e., knowledge transferring to its corresponding feature), which was carefully recorded by the human assessors.

Table V
OVERALL CGL PERFORMANCE ON SLOT FILLING

Slot Filling Systems		Apply CGL	P (%)	R (%)	F (%)
Our system		before	17.1	30.9	22.0
		after (f1+f4)	26.2	27.2	**26.7**
		after (full)	38.5	24.7	**30.1**
Other systems	High-Performing	before	13.7	29.6	18.8
		after (f1+f4)	40.9	22.2	**28.8**
	Moderate-Performing	before	12.2	7.4	9.2
		after (f1+f4)	35.7	6.2	**10.5**
	Low-Performing	before	6.7	3.7	4.8
		after (f1+f4)	50	3.7	**6.9**

Table VI
COST AND CONTRIBUTION OF EACH COMMENT

Comments		baseline	1	2	3	4
Performance	P (%)	17.1	17.6	26.4	26.7	25.6
	R (%)	30.9	30.9	28.4	28.4	27.2
	F (%)	22.0	22.4	27.4	27.5	26.3
Cost	#samples reviewed	-	3	3	3	3
	providing comments (minutes)	-	3	3	3	3
	encoding comments (minutes)	-	30	240	60	30

Table VI indicates that every feature made contributions to precision improvement. Comment 1 (gazetteer-based filtering) only provided limited gains mainly because our own system already extensively used similar gazetteers for answer filtering. This reflects a drawback of our comment generation procedure - the assessor had no prior knowledge about the approaches used in the systems. Comment 2 (using coreference resolution to check sentence occurrence) took most time to encode but also provides significant improvement. Comment 4 (consistency checking against responses with high votes) provided significant gains in

 ISBN: 978-1-61208-162-5

precision (8.5%) but also some loss in recall (3.7%). The problem was that systems tend to make similar mistakes, and the human assessor was biased by those correct answers which appeared frequently in the pooled system output. However, Comment 4 was able to filter out many errors which are otherwise very difficult to detect. For example, because "Najaf" appears very often as a "cities_of_residence" in the pooled system responses, Comment 4 successfully removed six incorrect "countries_of_residence" answers for the same query: "Syrian", "Britain", "Iranian", "North Korea", "Saudi Arabia" and "United States". On the other hand, Comment 4 confirmed correct answers such as "New York" from "Brooklyn", "Texas" from "Dallas", "California" and "US" from "Los Angeles".

G. Impact of Data Size

We also did a series of runs to examine how our own system performed with different amounts of training data. These experiments are summarized in Figure 2. It clearly shows that the learning curve converges quickly. Therefore, we only need a very small amount of training data (36 samples, 20% of total) in order to obtain similar gains (6.8%) as using the whole training set.

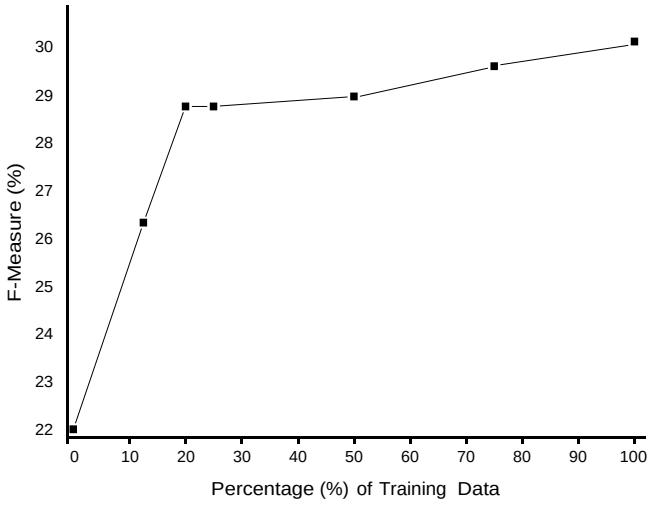


Figure 2. Impact of Training Data Size

H. Speed up Human Assessment

Human assessment for slot filling is also a costly task because it requires the annotators to judge each answer against the associated source document. Since our CGL approach achieved positive impact on system output, can it be used to as feedback to speed up human assessment? We applied the CGL assessor trained from comment 1 and comment 4 to the top 13 KBP systems for KBP2010 evaluation set. We automatically ranked the pooled system responses of residence slots according to their confidence values from high to low.

For comparison, we also exploited the following methods:

- Baseline
  As a baseline, we ranked the responses according to the alphabetical order of slot type, query ID, query name and answer string and doc ID. This is the same approach used by LDC human annotators for assessing KBP2010 system responses.
- Oracle (Upper-Bound)
  We used an oracle (for upper-bound analysis) by always assessing all correct answers first.

Figure 3 summarizes the results from the above 4 approaches. For this figure, we assume a labor cost for assessment proportional to the number of non-NIL items assessed. Note that all redundant answers are also included in these counts because human assessors also spent time on assessing them. This is only approximately correct; it may be faster (per response) to assess more responses to the same slot. The common end point of curves represents the cost and benefit of assessing all system responses. We can see that if we employ the CGL assessor and apply some cut-off, the process can be dramatically more efficient than the regular baseline based on alphabetical order. For example, in order to get 79 correct answers (76% of total), CGL approach took human assessors only 5.5 hours, while the baseline approach took 13.4 hours.

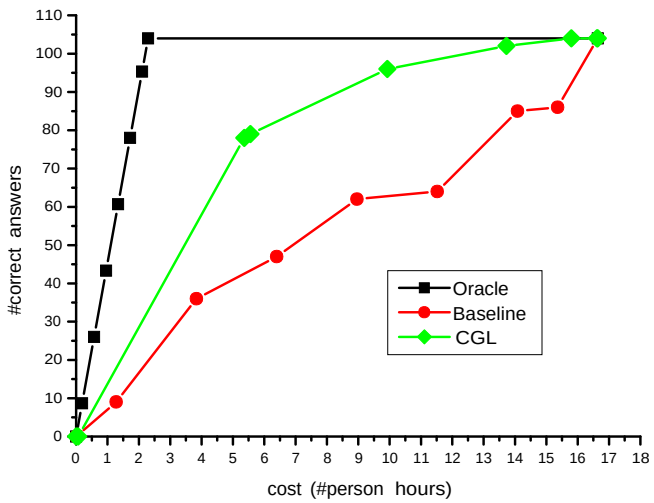


Figure 3. Human Assessment Method Comparison

I. Discussion

An alternative approach to validate answers is to use textual entailment techniques as in the RTE-KBP2010 validation pilot study [4], which was partly inspired by CLEF Question Answering task [19]. This task consists of determining whether a candidate answer (hypothesis "H") is supported in the associated source document (text "T") using entailment techniques. For the residence slots, we are considering in this paper, they treat each context document as a "T",

 ISBN: 978-1-61208-162-5

and apply pre-defined sentence templates such as "[Query] lived in [Answer]" to compose a "H" from system output. Entailment and reasoning methods from the TAC-RTE2010 systems are then applied to validate whether "H" is true or false according to "T". These RTE-KBP systems are limited to individual H-T instances and optimized only on a subset of the pooled system responses. As a result, they aggressively filtered many correct answers and did not provide improvement on most slot filling systems (including the representative ones we used for our experiment). In contrast, our CGL approach has the advantage of exploiting the generalized knowledge and feedback from assessors across all queries and systems.

VI. Conclusion and Future Work

To sum up, we have described a new validation approach called comment-guided learning (CGL). We demonstrated the power and generality of this approach on two very different applications: name translation and slot filling. Our approach significantly improved the performance of system responses and speeded up human assessment. It also outperformed some traditional validation methods, which, unlike ours, involved a great deal of feature engineering effort. The novelty of our approach lies in its declarative use of assessment feedback which may address some typical errors that a system tends to make. Some of such feedback will be otherwise difficult to acquire for feature encoding (e.g., Comment 3 in name translation and Comment 4 in slot filling). On the other hand, the simplicity of our approach lies in its low cost because it incorporates the bi-product of human assessment, namely their comments and explanations, instead of tedious instance-based human correction into the learning process. In this way the human assessor's knowledge is naturally transferred to the automatic assessor. Hence, CGL is amenable to implement but pertinent to a series of common errors identified.

In the future, we are interested in extending this idea to improve other NLP applications and integrating it with human reasoning. The current setup mainly improved precision but we also plan to embrace the idea of revertible query in question answering literature (e.g., [20]) and relation graph traverse to enhance recall. Ultimately we intend to investigate automatic ways to prioritize comments and convert comments to features so that we can better simulate the role of teacher in human learning.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER grant No. 1144111 and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] V. Vapnik, "Learning with Teacher: Learning Using Hidden Information," in Proc. International Joint Conference on Neural Networks, 2009.

[2] R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu "Human Active Learning," in Proc. NIPS2008, 2008.

[3] E. Bril, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," in Computational Linguistics, vol. 21, no. 1, 1995.

[4] K. Williams, C. Dozier, and A. McCulloh, "Learning transformation rules for semantic role labeling," in Proc. CoNLL-2004, 2004, pp. 134-137.

[5] Y. Al-Onaizan and K. Knight, "Translating named entities using monolingual and bilingual resources," in Proc. ACL2002, 2002, pp. 400-408.

[6] L. Bentivogli, P. Clark, I. Dagan, H. Dang, and D. Giampiccolo, "The sixth pascal recognizing textual entailment challenge," in Proc. TAC 2010 Workshop, 2010.

[7] Z. Chen, S. Tamang, A. Lee, X. Li, W.-P. Lin, M. Snover, J. Artiles, M. Passantino, and H. Ji, "Cuny-blender tackbp2010 entity linking and slot filling system description," in Proc. TAC 2010 Workshop, 2010.

[8] V. Castelli, R. Florian, and D. Han, "Slot filling through statistical processing and inference rules," in Proc. TAC 2010 Workshop, 2010.

[9] G. Bouma, S. Duarte, and Z. Islam, "Cross-lingual Alignment and Completion of Wikipedia Templates," in Proc. the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, 2009, pp. 21-29.

[10] D. Lin, S. Zhao, B. V. Durme, and M. Pasca, "Mining Parenthetical Translations from the Web by Word Alignment," in Proc. ACL2008, 2008, pp. 994-1002.

[11] G. You, S. Hwang, Y. Song, L. Jiang, and Z. Nie, "Mining Name Translations from Entity Graph Mapping," in Proc EMNLP2010, 2010, pp. 430-439.

[12] W.-P. Lin, M. Snover, and H. Ji, "Unsupervised language-independent name translation mining from wikipedia infoboxes," in Proc. EMNLP2011 Workshop on Unsupervised Learning for NLP, 2011, pp. 43-52.

[13] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in Soviet Physics Doklady, 1966.

[14] H. Ji, D. Westbrook, and R. Grishman, "Using semantic relations to refine coreference decisions," in Proc. HLT/EMNLP 05, 2005, pp. 17-24.

 ISBN: 978-1-61208-162-5

[15] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the tac 2010 knowledge base population track," in Proc. TAC 2010 Workshop, 2010.

[16] M. Richardson and P. Domingos, "Markov logic networks," in Machine Learning, 2006.

[17] K. Bollacker, R. Cook, and P. Tufts, "Freebase: A shared database of structured general human knowledge," in Proc. National Conference on Artificial Intelligence (Volume 2), 2007.

[18] R. Grishman, D. Westbrook, and A. Meyers, "Nyu's english ace 2005 system description," in Proc. ACE2005, 2005.

[19] A. Penas, A. Rodrigo, V. Sama, and F. Verdejo, "Testing the reasoning for question answering validation," in Journal of Logic and Computation, 2007.

[20] J. Prager, P. Duboue, and J. Chu-Carrol, "Improving qa accuracy by question inversion," in Proc. ACL-COLING2006, 2006, pp. 1073-1080.

 ISBN: 978-1-61208-162-5

Data Preparation in the *MineCor* KDD Framework

Christian Ernst
Ecole des Mines de St Etienne
CMP - Site Georges Charpak
Gardanne, France
ernst@emse.fr

Alain Casali
Laboratoire d'Informatique Fondamentale de Marseille (LIF),
CNRS UMR 6166, Aix Marseille Universités, IUT d'Aix en Provence
Aix en Provence, France
alain.casali@lif.univ-mrs.fr

***Abstract*—Yield enhancement is a key issue in semiconductor manufacturing. Data mining tools can therefore be helpful, by extracting hidden links between numerous complex process control parameters. In order to highlight correlations between such parameters, we developed a complete Knowledge Discovery in Databases (KDD) model, called *MineCor*. Its mining heart uses a new method derived from association rules programming, based on lectic search and contingency vectors. After recalling these concepts, this paper focuses on data preprocessing and transformation functions, which have an important impact on final results. An overall presentation of these functions, of some significant experimental results and of associated performances are provided and finally discussed.**

***Keywords*-Data Mining, Semiconductor Manufacturing, Decision Correlation Rule, Data Preparation.**

I. Introduction and motivation

In this Section, we first introduce why and how data mining techniques are useful to detect the main parameters, which have an impact on yield loss in semiconductor fabrication capabilities. Then, we present our approach, based on a complete KDD model. It determines the main correlated production parameters impacting the yield, and is based on important preliminary preparation tasks.

A. Data mining techniques in semiconductor fabs

Data Mining specifies data models, which may be rules, anomalies or trends that are of interest. To improve quality in manufacturing areas, mining techniques extract knowledge to identify hidden patterns in the parameters that control production processes [1]. Unfortunately, there are no scalable models for associated applications, but only "implementation specific" mining algorithms. We focus hereafter on (i) Fault detection and quality improvement, which examine what happened in the past to better predict and to improve the future system's performance; and (ii) Decision support systems, which determine links between control parameters and product quality in the form of rules. We moreover concentrate on a particular area, semiconductor wafer manufacturing, where yield is the ratio of non-defective chips in a finished wafer to the number of input products.

In semiconductor manufacturing facilities, the volume and the complexity of the collected data are generally much more consequent than in other manufacturing fields: Fabrication processes include several hundred steps with regard to the produced chip. Each step uses various chemico-physical recipes, grouped into four phase units (photolithography, etch, implant and CMP).

Two techniques are used to improve the yield: Real-time and *post hoc*. The first approach monitors on-line measurements of process steps, and undertakes corrective action to ensure that the measures remain within desired limits. The *post hoc* approach compares the end result of the whole process with the desired specifications, analyzing the root causes of low yield for adjusting the process parameters to ensure future quality. Advanced Process Control (APC) considers both, by highlighting correlations between production parameters in order to rectify possible drifts of the associated process(es). This can be done for specific equipment and process steps in real-time: FDC (Fault Detection and Classification) tools and R2R (Run To Run) regulation loops are the most representative APC techniques. Correlations can also be discovered *post hoc*: This is the framework of our paper.

Both approaches first try to identify, which parameters are the root causes of a particular yield excursion. However, conventional methods such as SPC are here inaccurate, because they fail to extract underlying features from complex data. This is why data mining techniques become useful in semiconductor fabs. Associated models can be categorized into four types [2]: Classification, Clustering, Prediction and Association Rules. The most widely used is classification. At the contrary, association rules are not often used. Let us mention [3], where the authors present a modified *Apriori* algorithm used in LCD panel manufacturing in order to locate machines with low yield after process completion.

B. Our approach

We present a whole KDD model based on specific rules. Within this framework, and in collaboration with STMicroelectronics and ATMEL, our work is focused on the detection of the main control parameters impacting the yield. The goal is to propose indicators to which special attention should be paid in order to construct, in a second step, yield

 ISBN: 978-1-61208-162-5

enhancement models used in further production cycles. Let us emphasize that this second non-trivial problematic is excluded from the scope of our paper.

The realized *post hoc* analysis is based on CSV files of real valued measurements associated with production lots, extracted from large databases and covering the four fabrication units mentioned above. The main characteristic of these files is the huge number of columns (nature of the measurements) with regard to the number of rows (measures). We want to highlight correlations between the values of some columns and those of a target column: The yield. To detect these correlations, we use Decision Correlation Rules [4] but, before, we undertake important data preparation tasks to enhance the efficiency of the mining input database: This is the scope of our paper.

The paper is organized as follows: In Section II, the bases of Decision Correlation Rules and our mining algorithm are first recalled. In Section III, we expose the data preparation functions of the *MineCor* software. Experiments are detailed in Section IV. As a conclusion, we summarize our contributions and outline some research perspectives.

II. Related work:The MineCor mining model

In this section, we first recall the definitions of correlation rules and lectic order. Then we introduce the Ls Algorithm, which allows to browse the search space according to the lectic order, before presenting the LHS-Chi2 algorithm. Some points developed hereafter have soon been presented in [4]. But the given overview clarifies the approach.

A. Decision correlation rules and lectic order

Basic concepts

An association rule [5] is an approximate implication $X \rightarrow Y$ between two sets of items. Two measures are used to extract significant rules: Support and confidence. Because the underlying semantics of an association rule are fairly poor, Wu et al. [6] introduce literalsets and compute positive and/or negative association rules such as $\neg X \rightarrow Y$. To generate the rules, the authors use the same platform by redefining the support of a literal: The number of transactions of the binary relation including X and containing no 1-item of Y. Another approach is proposed by Brin *et al.* [7]: The extraction of correlation rules. The new platform is based on the Chi-Squared statistical measure, written χ^2. We assume hereafter that the definitions of literalsets so as of the χ^2 statistic are known.

When computing correlation rules, the memory usage required by levelwise algorithms is crucial. This is why Brin *et al.* compute only correlations between two values. Different criteria to evaluate whether a correlation rule is semantically correct have been proposed. The main is the Cochran criterion, and can be relaxed as follows: $MinPerc$ of the literalsets of a contingency table must have a support larger than $MinSup$, where $MinPerc$ (minimal percent) and $MinSup$ (minimal support) are thresholds.

Decision correlation rules

We want the computed correlation rules to include specific items, *e.g.* belonging to a target attribute. Let r be a binary relation (a transaction database) over a set of items $\mathcal{R} = \mathcal{I} \cup T$. $\mathcal{I}$ represents the values (the items) of the binary relation used as analysis criteria, and $\mathcal{T}$ is a target attribute, which items may be null.

Definition 1 (Decision Correlation Rules): Let $X \subseteq \mathcal{R}$ be a pattern, and $MinCorr$ a given threshold (≥ 0). X represents a valid Decision Correlation Rule if and only if: (i) X contains a value of the target attribute $\mathcal{T}$; and (ii) $\chi^2(X) \geq MinCorr$.

Table I
Relation example r.

Tid	ItemSet	Target
1	BCF	t_1
2	BCE	t_1
3	BCF	t_2
4	BC	-
5	BD	t_1
6	B	-
7	ACF	t_1
8	AC	-
9	AE	t_1
10	F	t_2

Example 1: With the relation Example r given in Table I, Table II shows the contingency table of pattern BC.

Table II
Contingency table of pattern BC.

	B	$\overline{B}$	$\sum_{row}$
C	4	2	6
$\overline{C}$	2	2	4
$\sum_{column}$	6	4	10

Continuing the example, $\chi^2(BC) \simeq 0.28$. which corresponds to a correlation rate of about 45%. If $MinCorr = 0.25$, the correlation rule materialized by the BC pattern is valid, but the correlation rule represented by the Bt_1 pattern is not ($\chi^2(Bt_1) \simeq 0.1$).

Lectic order and Lectic subset algorithm

The lectic order [8], noted $<_{lec}$, permits to enumerate all the subsets of an itemset $\mathcal{I}$.

Definition 2 (Lectic Order): Let $\mathcal{I}$ be a set of items totally ordered and therefore comparable two by two via an order denoted by $\preceq$. If X and $Y \subseteq \mathcal{I}$, then we have: $X <_{lec} Y \Leftrightarrow max_{\preceq}(X \backslash (X \cap Y)) \preceq max_{\preceq}(Y \backslash (X \cap Y))$.

The Lectic Subset Algorithm, noted Ls [9], is one of its possible implementations.

 ISBN: 978-1-61208-162-5

B. The LHS-Chi2 Algorithm

Equivalence classes

In [4], we adapted the concept of equivalence classes over literal patterns to our context:

Definition 3 (Equivalence Class associated with a literal): Let $Y\overline{Z}$ be a literal, and $[Y\overline{Z}]$ its associated equivalence class. This class contains the set of transaction identifiers of the relation including Y and containing no value of Z.

Example 2: With our relation Example (see Table 1), we have $[B\overline{C}] = \{5, 6\}$.

Contingency vectors

The contingency vectors are another representation of the contingency tables:

Definition 4 (Contingency Vector): Let $\mathbb{P}(X)$ be the powerset lattice of X, and $X \subseteq \mathcal{R}$ a pattern. The contingency vector of X, denoted $CV(X)$, groups the set of the literalset equivalence classes belonging to $\mathbb{P}(X)$ ordered according to the lectic order.

Example 3: With our sample relation (see Table 1), we obtain $CV(BC) = \{[\overline{BC}], [B\overline{C}], [C\overline{B}], [BC]\} = \{\{9, 10\}, \{5, 6\}, \{7, 8\}, \{1, 2, 3, 4\}\}$.

Proposition 1 is the main result presented in [4]. It shows how to compute the CV of the $X \cup A$ pattern given the CV of X and a set of identifiers of the relation containing A.

Proposition 1: Let $X \subseteq \mathcal{R}$ be a pattern and $A \in \mathcal{R} \backslash X$ a 1-item. The CV of the $X \cup A$ pattern can be computed given the CVs of X and A as follows:

$$CV(X \cup A) = (CV(X) \cap [\overline{A}]) \cup (CV(X) \cap [A])$$

Example 4: With the relation Example (see Table 1), we have $CV(B) = \{\{7, 8, 9, 10\}, \{1, 2, 3, 4, 5, 6\}\}$ and $CV(C) = \{\{5, 6, 9, 10\}, \{1, 2, 3, 4, 7, 8\}\}$. By applying Proposition 1 and ordering, we retrieve the result of Example 3: $CV(BC) = \{\{9, 10\}, \{5, 6\}, \{7, 8\}, \{1, 2, 3, 4\}\}$.

The LHS-CHI2 Algorithm

The Lectic Hybrid Subset-Chi2 Algorithm, or LHS-CHI2, adapts the LS Algorithm to our context, in the way that it includes contingency vectors so as five constraints in order to prune the search space [4]. The predicate *CtPerc* expresses the satisfiability of the Cochran criterion. The pseudo-code of the procedure CREATE_CV can be found in [4]. By convention, we consider that we have $CV(\emptyset) = \{Tid(R),\ \emptyset\}$. The positive border ($BD^+$) is initialized with $\{\emptyset\}$. The pseudo code of LHS-CHI2 is provided in Algorithm 1. The first call to LHS-CHI2 is carried out with $X = \emptyset$ and $Y = \mathcal{R}$.

Example 5: The Decision Correlation Rules computed by LHS-CHI2 on the Example of Table 1 and satisfying the minimal threshold constraints $MinSup = 0.2$, $MinPerc = 0.25$ and $MinCorr = 0.25$ are shown in Table III.

Algorithm 1: LHS-CHI2 Algorithm.

input : X and Y two patterns
output: $\{\ Z \subseteq X$ such that $\chi^2(Z) \geq MinCorr\}$

```
if Y = ∅ and ∃t ∈ T : t ∈ X and |X| ≥ 2 and χ²(X) ≥ MinCorr then
    Output X, χ²(X)
end
A := max(Y) ;
Y := Y\{A} ;
LHS-CHI2(X,Y) ;
Z := X ∪ {A} ;
if ∀z ∈ Z, ∃W ∈ BD⁺ : {Z\z} ⊆ W then
    CV(Z) := CREATE_CV(VC(X),Tid(A)) ;
    if |Z| ≤ MaxCard and CtPerc(CV(Z), MinPerc, MinSup) then
        BD⁺ := max⊆(BD⁺ ∪ Z) ;
        LHS-CHI2(Z,Y) ;
    end
end
```

Table III
RESULTS OF THE LHS-CHI2 ALGORITHM OVER TABLEI

Decision Correlation Rule	χ^2 Value
At_1	0.48
BCt_1	0.28
BFt_1	0.28

Performance issues

We show in [4] that levelwise algorithms require to store CTs 2.5 GB of memory at the 3rd level, and 1.3 TB at the 4th level. When our algorithm requires, in the worst case, $2|r| * (MaxCard + n + 1)$ Bytes in memory, where n is the number of 1-items in the relation r. Which is much less than the mentioned volumes above. This is because we need only to keep the CVs stored in each branch of the search tree, the computation of a CV at each level using the CVs memorized at the upper levels. When a branch has been pruned, all the stored CVs in that branch are released. Moreover, computing a CV is faster than computing a CT.

III. DATA PREPARATION WITHIN *MineCor*

We developed a global KDD model including the LHS-Chi2 algorithm. The software, called *MineCor* (*Mine*r for *Cor*relations), is developed in C language. To carry out preprocessing and transformation in the form of a transaction database of the input files, we first performed column elimination and discretization stages [2], [10]. Description of these steps are our main contribution: seldom discussed nor presented in an industrial context, they have a huge impact on final results, and are summarized in Sections III-A and III-B. The output of the two steps is the source for the data

mining phase. Which is followed by an interpretation of the results, resumed in Section III-C.

A. Preprocessing stage

The first step of data cleaning is the preprocessing stage. Preprocessing consists in the reduction of the data structure [11] by eliminating columns and rows of low significance. This is for two reasons: (i) If each value of each column is considered as a single item, there will be a combinatorial explosion of the search space, and thus very large reponse times; (ii) We cannot expect this task to be performed by an expert, because manual cleaning of data is time consuming and subject to many errors. The defined order of following functions is important.

Elimination of Concentrated Data and Outliers

We eliminate first columns having small standard deviation (threshold $MinStd$): Since the values are almost the same, we consider that they do not have a significant impact on the result; but their presence pollutes the search space and reduces the response times. In the same way, we introduced the $MinDV$ threshold, which imposes a Minimal number of Distinct Values in each column. Attention is finally paid to inconsistent values, such as "outliers" in noisy columns. Detection is performed through another convenient threshold ($fStd$, a factor of $MinStd$), and elimination consists to force the detected values to null.

Other Column Elimination

The dysfunction of sensors, or the occurrence of a maintenance step may imply that some sensors can not transmit their values to the database. As a consequence, the associated columns contain many null/default values, and are thus deleted from the input file. Such cases are here detected using the $MaxNV$ (Maximum Null Values) threshold. Moreover, sometimes, several sensors measure the same information, resulting in identical columns in the source file. In this case, we keep only a single column. Finally, columns with no item having the support ($MinSup$ threshold) are also removed.

Normalization

Let $\mathcal{S}$ be the set of values to be discretized (the input column values as a numeric vector), and $Min_{\mathcal{S}}$ and $Max_{\mathcal{S}}$ be the smallest and the largest value of $\mathcal{S}$. Finally, and in order to manage the different values associated with each set S in the same way, we normalize the values to keep them between 0 (kZero) and 1 (kOne). This is performed by replacing each value $v \in \mathcal{S}$ by $\frac{v-Min_{\mathcal{S}}}{Max_{\mathcal{S}}-Min_{\mathcal{S}}}$.

B. Discretization stage

Discrete values deal with intervals of values, which are more concise to represent knowledge, so that they are easier to use and comprehend than continuous values. Many discretization algorithms have been proposed over the years in order to classify data into intervals, also called bins. Discretization can thus be performed [12]:

- In a supervised or unsupervised manner, depending on whether class information is at one's disposal;
- In a dynamic or static way: With a static approach, discretization is done before the classification task;
- Using splitting or merging techniques: In the latter case, the search space is examined bottom-up.

We represent continuous real valued columns by associating to each of their values an interval code (the one to which the value belongs). The intervals are created using either equal-width, equal-frequency and embedded means discretization, which are non supervised, static and splitting methods. In each approach, NIC is an input parameter specifying the number of bins to create per column.

Equal Width Discretization (EWD)

Each interval has a length of $l = \frac{Max_{\mathcal{S}}-Min_{\mathcal{S}}}{NIC}$. The computed classes are $c_1 : [Min_{\mathcal{S}}, Min_{\mathcal{S}} + l[$, $c_2 : [Min_{\mathcal{S}} + l, Min_{\mathcal{S}} + 2l[$, ... The method is easy to compute and to interpret, but is not efficient in the case of asymmetric or discontinuous distributions.

Equal Frequency Discretization (EFD)

The goal is to obtain classes having, if possible, the same number of values. Because this configuration seldom appears, the problem becomes how to group close values into classes while respecting the above constraint. The Jenks' natural breaks classification schema gets the best class arrangement, after having generated each possible class combination [13]. It minimizes the in-class difference and maximizes the between-class difference using the Goodness of Variance Fit (GVF):

$$GVF = 1 - \frac{\sum_{j=1}^{NIC} \sum_{i=1}^{|[\mathcal{S}_i,\mathcal{S}_j]|} (\mathcal{S}_i - \overline{[\mathcal{S}_i,\mathcal{S}_j]})^2}{\sum_{i=1}^{|[\mathcal{S}]|} (S_i - \overline{\mathcal{S}})^2},$$

where $|[\mathcal{S}_i,\mathcal{S}_j]|$ is the cardinality of the interval $[\mathcal{S}_i,\mathcal{S}_j]$, and $\overline{\mathcal{S}}$ is the mean of the sorted set $\mathcal{S}$.

The main drawbacks of the method are, on one hand, that stability is not assumed when NIC varies and, on the other hand, the high computational complexity of the class generation, which is C_{d-1}^{NIC-1}, where d is the number of distinct values in the set $\mathcal{S}$. The associated computational cost becoming redhibitory, we use instead the Fisher's method of exact optimization [14] proposed for grouping $|[\mathcal{S}]|$ elements (distinct or not) into NIC mutually exclusive and exhaustive subsets having maximum homogeneity, *i.e.,* minimizing the within-groups sum of squares. The obtained partition is guaranteed to be optimal, but not unique. Which is not important while the obtained gain of time is.

 ISBN: 978-1-61208-162-5

Embedded Means Discretization (EMD)
EMD is a divisive hierarchical clustering method, which starts with a single class/bin that contains all the initial values of the column to discretize ($\mathcal{S}$). The average divides the set into two groups used to construct two new classes. The averages of these 2 groups permit the splitting into 4 classes, and so on. This approach matches every kind of distribution, but the number of classes NIC is here always a power of two, which may be an inconvenient.

Algorithm 2: REM Algorithm.

input : Vec : initial vector, Min, Max 2 borders, ind : current index, nbi : number of bins
output: $MVec$: average vector

$Avg := ComputeAvg(Min, Max, Vec)$;
$MVec[Ind] := Avg$;
if $NIC > 1$ **then**
  REM $(Vec, Min, Avg, ind - nbi/2, nbi/2)$;
  REM $(Vec, Avg, Max, ind + nbi/2, nbi/2)$;
end

We compute the class borders ($MVec$) using the Recursive Embedded Means (REM) discretization algorithm, which pseudo code is provided in Algorithm 2. The first recursive call to REM is carried out with $Vec = \mathcal{S}$, $Min = kZero$, $Max = kOne$, $ind = NIC/2$ and $nbi = NIC/2$. The $ComputeAvg$ function returns the average of the Vec values bordered by Min and Max.

String Valued Columns Discretization
Because string valued columns often appear as parameter measures, we take them equally into account by applying them a rough discretization function presented in Algorithm 3. Associated Compute String Intervals (CSI) discretization method keeps the NIC or $NIC - 1$ most present string literals of input string vector $\mathcal{S}$ as discretization "values". If the $bOth$ boolean is enabled, $OutVec[NIC - 1]$ represents any other value of $\mathcal{S}$ not equal to any of the first $NIC - 1$ values of $OutVec$. The column is removed for the mining step if it contains strictly less than NIC values.

C. Interpretation stage

Interpretation essentially consists in decoding the discretization stage with regard to the results, and to produce an intelligible output for the end-user. *MineCor* produces outputs in HTML and text formats.

Example 6: BCt_1 is a valid Decision Correlation Rule (cf Table III). Associated text output looks like [1.4, 2]; [2.8, 4.6]; [2.7, 7.4], where $[b_{min}, b_{max}]$ are real values representing items B, C, t_1 respectively.

IV. Experimental Analysis

Some representative results of the LHS-Chi2 algorithm are presented below. As emphasized in Section I-B, the

Algorithm 3: CSI Algorithm.

input : nbi : number of bins , $InVec$: input string vector, $bOth$: "others" boolean, $Minsup$: threshold
output: $OutVec$: output string vector

$OutVec := \emptyset$;
$nbDisVals := SortByPopularity(InVec, OutVec)$;
if $nbDisVals \geq NIC$ *and* $bOth$ *and* $Sup(OutVec[nbi - 1]) \leq MinSup$ **then**
  $OutVec[nbi - 1] := kOthers$;
end

experiments were done on different CSV files of (essentially) real value measures supplied by STMicroelectronics (STM) and ATMEL (ATM). The files have one or more target columns, resulting from the concatenation of several measurement files. The characteristics of the datasets used can be found in Table 4. STM File1 and ATM File are representative of our *post hoc* approach (*cf.* Section I-A). STM File2 is more typical of a real-time approach (few parameters, large number of measures).

Table IV
Dataset Examples

Name	Number of Columns	Number of Rows
STM File1	1 281	297
STM File2	8	726
ATM File	749	213

All experiments were conducted on an HP Workstation (1.8 GHz processor with a 4 Gb RAM). Results are presented on Figures 1 through 7(b).

In [4], we compare the execution times of a classical Levelwise algorithm and LHS-Chi2. The experiments use the only EWD method. The response times of our method are between 30% and 70% better than Levelwise.

A. Impact of the Preprocessing Stage

Figures 1 and 2 show respectively the number of items used in the mining stage and the number of decision correlation rules discovered after that stage for STM File2 (target1) when $MinSup$ (0.2), $MinCorr$ (0.24) and $MinPerc$ (1.2) are constant, while the number of bins (NIC) varies. We compare the three discretization methods.

EMD (and EFD) are better methods than EWD when the NIC parameter remains low. With greater values of NIC and also because of the large number of rows in the analyzed file, EWD returns more rules (even if not necessary all of interest). As it appears also on the following experiments, EWD is the best method the larger the thresholds, because of the more important number of columns kept and thus of

 ISBN: 978-1-61208-162-5

Figure 1. Number of items used *vs*. number of bins.

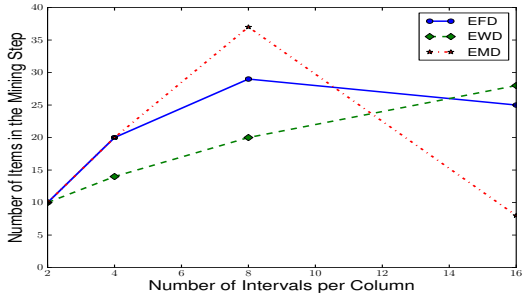
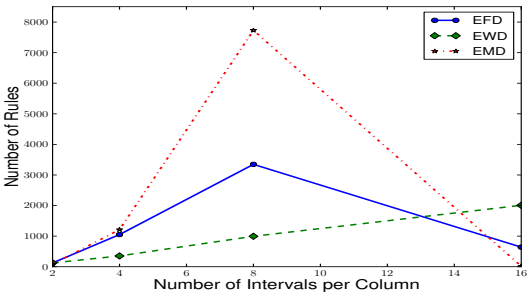


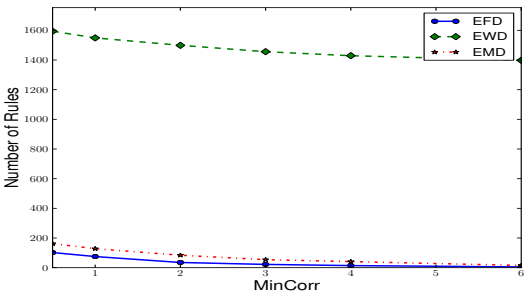
Figure 2. Number of generated rules *vs*. number of bins.



items produced after preprocessing, even if the execution times are more important.

Figures 3 and 4 show the number of extracted rules after mining when $MinPerc$ (0.34) and $MinSup$ (0.25 in Figure 3 and 0.27 in Figure 4) are fixed. The difference between the two experiments is that the $MinDV$, $MaxNV$ and $MinStd$ thresholds are smaller in Figure 4 than in Figure 3. What harmonizes the results, and shows also that the $MinCorr$ threshold has only few effect on mining. What means also that the preprocessing parameters impact the numbers of obtained items and thus of computed rules.

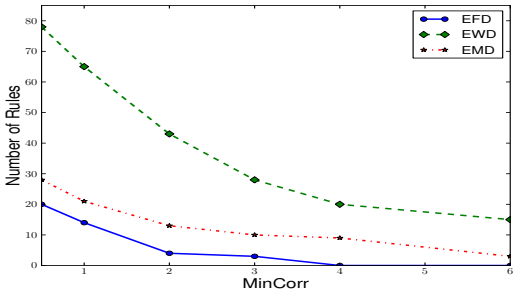
Figure 3. Number of generated rules with $MinDV = 58$, $MaxNV = 48$, $MinStd = 1.68$ (STM File1 - target1).



B. Impact of the Discretization Stage

Figures 5(a) and 6(a) show the number of items kept after the discretization stage, which only depends on the $MinSup$ threshold, while the number of bins is constant. They illustrate that the smaller the threshold $MinSup$, the larger the number of items kept for the mining stage, whatever the discretization method. Figures 5(b) and 6(b) show the number of rules that are generated in both cases. While the number of partitions generated by the EFD method is larger than the one generated by the EWD method, the number of rules is smaller. Moreover, the execution time is shorter by a factor up to 2.5 (*cf.* Figure 5(c)). On the other hand, the EMD method provides better results when working on STM File2, which is a perticular case. These results outline that *MineCor* tries to provide the end-user with "best" quality rules: (i) Low in number, (ii) Significant, and (iii) Computed quickly.

Figure 4. Number of generated rules with $MinDV = 80$, $MaxNV = 50$, $MinStd = 1.28$ (STM File1 - target1).



V. Conclusion and Future Work

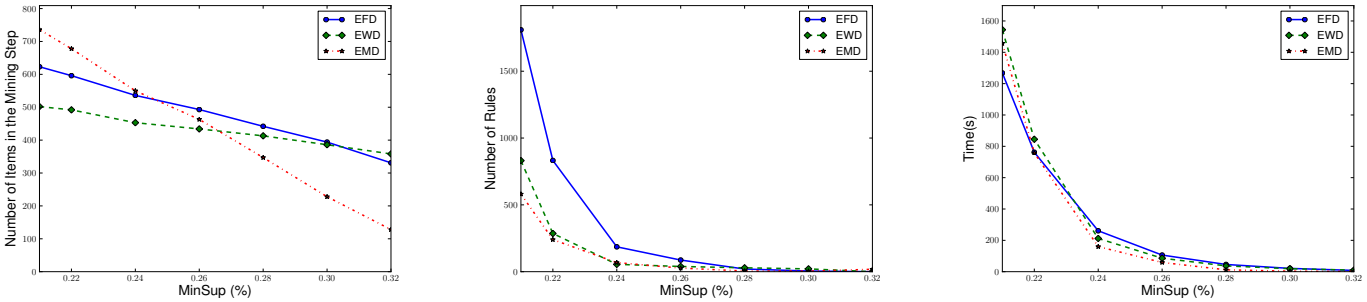
In this paper, we presented the *MineCor* software. Parameter measurement files given by semiconductor manufacturers are used, and produce values of parameters with most influence on the yield. To achieve this objective, we built a complete KDD model, based on Decision Correlation Rules. We show that the various thresholds used in our preprocessing stage have an impact on the number of kept columns of the input file, and thus, on the number of items used in further steps. Moreover, we implemented three methods at the discretization stage: (i) Equal Width, (ii) Equal Frequency, and (iii) Embedded Means. Experiments point out that, in most cases, the EFD method produces Decision Correlation Rules faster and of better quality.

Some new issues to our work are: (i) To compare our approach with classification methods; (ii) To optimize the processing stages upstream of the algorithm (aggregation of attributes, merging of intervals) while safeguarding the context in order to obtain a larger number of rules and/or more significant results; and (iii) To allow automatic threshold and parameter fixing depending on each input file column.

References

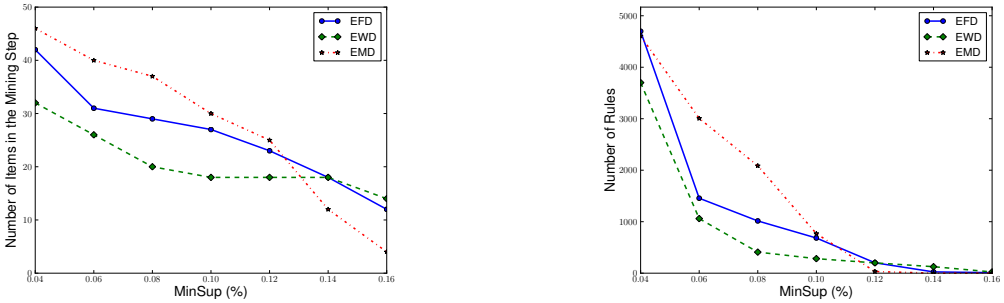
[1] A. Choudhary, J. Harding, and M. Tiwari, "Data mining in manufacturing: a review based on the kind of knowledge," *Journal of Intelligent Manufacturing*, vol. 20, no. 5, pp. 501–521, 2009.

Figure 5. Results with 4 intervals, $CtPerc = 0.34$, $MinCorr = 2.8$ (ATM File - target3).



(a) Number of items kept after the Preprocessing and Discretization stages (b) Number of generated Decision Correlation Rules (c) Execution Time

Figure 6. Results with 8 intervals, $CtPerc = 0.24$, $MinCorr = 1.6$ (STM File2 - target1).



(a) Number of items kept after the Preprocessing and Discretization stages

(b) Number of generated Decision Correlation Rules

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[3] C. Huang and R. Chen, "Application of new apriori algorithm mdnc to tft-lcd array manufacturing yield improvement," *International Journal of Computer Applications in Technology*, vol. 28, pp. 161–168, 2007.

[4] A. Casali and C. Ernst, "Extracting decision correlation rules," in *DEXA*, ser. Lecture Notes in Computer Science, S. S. Bhowmick, J. Küng, and R. Wagner, Eds., vol. 5690. Springer, 2009, pp. 689–703.

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996, pp. 307–328.

[6] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules," *ACM Trans. Inf. Syst.*, vol. 22, no. 3, pp. 381–405, 2004.

[7] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in *SIGMOD Conference*, 1997, pp. 265–276.

[8] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

[9] M. Laporte, N. Novelli, R. Cicchetti, and L. Lakhal, "Computing full and iceberg datacubes using partitions," in *ISMIS*, 2002, pp. 244–254.

[10] D. Pyle, *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

[11] O. Stepankova, P. Aubrecht, Z. Kouba, and P. Miksovsky, "Preprocessing for data mining and decision support," in *Data Mining and Decision Support: Integration and Collaboration*, K. A. Publishers, Ed., 2003, pp. 107–117.

[12] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," *Data Min. Knowl. Discov.*, vol. 6, no. 4, pp. 393–423, 2002.

[13] G. Jenks, "The data model concept in statistical mapping," in *International Yearbook of Cartography*, vol. 7, 1967, pp. 186–190.

[14] W. Fisher, "On grouping for maximum homogeneity," in *Journal of the American Statistical Association*, vol. 53, 1958, pp. 789–798.

 ISBN: 978-1-61208-162-5

Exploiting Background Information Networks to Enhance Bilingual Event Extraction Through Topic Modeling

Hao Li, Heng Ji
Computer Science Department
Queens College and Graduate Center, CUNY
New York, USA
{*haoli.qc,hengjicuny*}*@gmail.com*

Hongbo Deng, Jiawei Han
Computer Science Department
University of Illinois at Urbana-Champaign
Urbana-Champaign, USA
{*hbdeng,hanj*}*@uiuc.edu*

***Abstract*—In this paper, we describe a novel approach of biased propagation based topic modeling to exploit global background knowledge for enhancing both the quality and portability of event extraction on unstructured data. The distributions of event triggers and arguments in topically-related documents are much more focused than those in a heterogeneous corpus. Based on this intuition, we apply topic modeling to automatically select training documents for annotation, and demonstrate it can significantly reduce annotation cost in order to achieve comparable performance for two different languages and two different genres. In addition, we conduct cross-document inference within each topic cluster and show that our approach advances state-of-the-art.**

Keywords-Event Extraction; Background Information Network; Biased Propagation based Topic Modeling.

I. INTRODUCTION

Event extraction is the task of identifying events of a particular type and their participants (arguments) from documents. It is a complex task which suffers from two major problems: (1) quality: challenges in disambiguating event types indicated by trigger words and roles played by arguments; (2) portability: high-cost of manual annotation in obtaining training data. With the rapid growth of new genres, such as web blogs and Twitter which is far more informal and noisy, these challenges become more critical. We found that event extraction performed notably worse on web blogs than on newswire texts. While labeled newswire documents are widely available, labeled informal texts are often expensive to obtain, and are generally scarcely available.

Most of the previous event extraction methods focused on improving the performance for one single document in isolation. When a typical event extraction system processes one document in a large collection, it makes only limited use of local 'facts' already extracted in the current document, such as names, noun phrases and time expressions. However, if we take one step back by looking at human learning, we often see that students study in groups of two or more, mutually searching for the best understanding, solution or meaning; researchers gather together as a "committee" or "panel" to select the best paper/project proposal. Such activities are formalized as "collaborative learning" [26]. Similarly, when dealing with large amounts of data, the event extraction task is naturally embedded in rich contexts. Events no longer exist on their own; they are connected to other topically-related documents, associated with authors (e.g., posters for blogs, reporters for news, speakers for conversation transcripts, authors for papers) and the publication venues (e.g., forums for blogs, agencies for news, conferences for papers), and linked to the geographical places where the documents are published. We call such heterogeneous contexts as the background "information networks" for each candidate event in a test document, as depicted in Figure 1. However, it is not trivial to encode such contextual clues directly into the event extraction system because they are ubiquitously interrelated in various network structures.

In this paper, we propose to directly incorporate multi-dimensional heterogeneous background information networks, through a new and uniformed biased propagation based topic modeling framework as described in our recent work [11].

The underlying intuition is that multi-typed contextual information should be integrated but treated differently in the topic model. This method is designed to imitate human collaborative learning to seek topically-related events as "collaborators" and enhance both training (portability) and testing (quality) an event extractor:

- *training*: automatically select topically-related documents as for training data annotation; we shall demonstrate that this method can significantly reduce annotation cost.
- *testing*: conduct statistical cross-document inference within each topic cluster to favor consistency of interpretation across documents and achieve higher extraction quality; we shall demonstrate topic modeling provides a more effective way than information retrieval (IR)-based document clustering.

We extensively evaluate the proposed approach and compare to state-of-the-art techniques on different data genres (newswire and web blogs) and different languages (English and Chinese). Experimental results demonstrate that the improvement in our proposed approach is language-independent, genre-independent, consistent and promising.

 ISBN: 978-1-61208-162-5

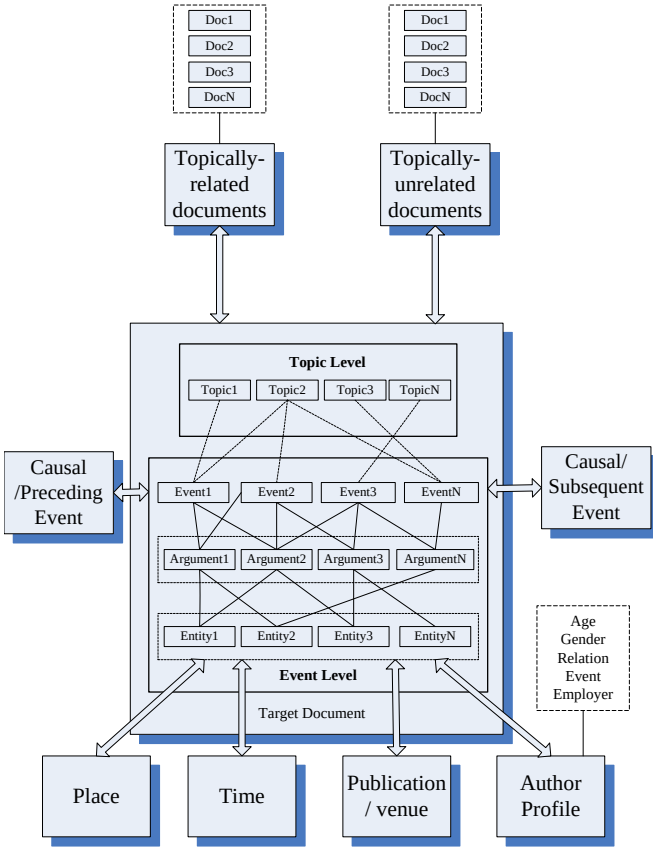


Figure 1. Background Information Networks for Event Extraction

The novel contributions of this paper are two-fold: (1) the first attempt to integrate background knowledge into topic modeling for general news and web blog domains; (2) the first work on exploiting topic modeling to improve both portability and quality of event extraction.

The paper is structured as follows. We briefly review related work in section 2. Section 3 introduces task definition and baseline systems. Then, we propose a novel topic modeling in section 4. In section 5, we apply the topic modeling on event extraction task. The experimental results are presented in section 6. The conclusion and summary is presented in section 7.

II. Related Work

Some recent work exploited global background knowledge to enhance information extraction tasks, such as entity coreference resolution [25], entity linking [9][13] and relation extraction [7]. Most of these methods incorporated background knowledge from external resources (e.g., Wikipedia). Several recent IE studies have stressed the benefits of using information redundancy on estimating the correctness of the information extraction output [12][18][24][31] or conducting cross-event reasoning [14][21]. We apply topic modeling to locate specific background documents more accurately.

Recently topic models have been successfully applied to various fields of natural language processing, such as Information Retrieval (e.g., [22][29]), Word Sense Disambiguation(WSD) [5], Person Name Disambiguation [27], Text Categorization [34] and Temporal Event Tracking [15]. When reading on-topic stories to understand the events that happened, people tend to segment such stories into various activities (or topics) [32]. Previous research also recognized the benefits of organizing information by events, such as topic detection and tracking [2]. However, very little work has used topic information as feedback to improve event extraction.

In addition, almost all of these previous applications utilized topic models during the test phase, while we demonstrate that topic models can also be used as an effective way to select training data for event extraction, and thus predict the extraction performance before annotating the whole training set. Agichtein and Cucerzan [1] described a language modeling approach to quantify the difficulty of entity extraction and relation extraction. Active learning methods have been applied to reduce annotation cost for information extraction (e.g., [19]). Patwardhan and Riloff [23] also demonstrated that selectively applying event patterns to relevant regions can improve MUC event extraction. Our experiments suggested that topical relatedness can serve as a potential metric to be integrated into other standard data selection criteria in active learning.

III. Event Extraction Task and Baseline System

A. Task Definition

The event extraction task we are addressing is that of the Automatic Content Extraction (ACE) [20].

ACE defines the following terminology:

- Event type: a particular event class
- Event trigger: the main word which most clearly expresses an event occurrence
- Event arguments: the mentions that are involved in an event (participants) with particular roles

The 2005 ACE evaluation had 8 types of events, with 33 subtypes; for the purpose of this paper, we will treat these simply as 33 distinct event types. For example, the sentence "*the US-led coalition troops are reportedly thrusting into the second Iraqi city of Basra.*" includes a "*Movement_Transport*" event that is indicated by a trigger word ('*thrusting*"), and a set of event arguments: the Artifact ("*troops*") and the destination ("*Basra*").

We define the following standards to determine the correctness of an event mention:

- A trigger is correctly labeled if its event type/subtype and offsets match a reference trigger.

 ISBN: 978-1-61208-162-5

- An argument is correctly labeled if its event type/subtype, offsets, and role match any of the reference argument mentions.

B. Baseline Bilingual Event Extraction

We use two state-of-the-art event extraction systems ([8][18]) as our baseline, one for English and the other for Chinese. The system combines pattern matching with a set of Maximum Entropy classifiers incorporating diverse lexical, syntactic, semantic and ontological knowledge. It takes raw documents as input and conducts some pre-processing steps. The texts are automatically annotated with word segmentation, part-of-speech tags, parsing structures, entities, time expressions, and relations. The annotated documents are then sent to the following classifiers: to distinguish events from non-events; to classify events by type and subtype; to distinguish arguments from non-arguments; to classify arguments by argument role; and given a trigger, an event type, and a set of arguments, to determine whether there is a reportable event mention. In addition, the Chinese system incorporates some language-specific features to address the problem of word segmentation and special noun phrase structures. Each component can produce reliable confidence values.

IV. Capture Background Knowledge Through Topic Modeling

In this section, we will describe a novel topic modeling approach to integrate background knowledge.

A. Probabilistic Latent Semantic Analysis

Many topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [16] and Latent Dirichlet Allocation (LDA) [4], have been proposed and shown to be useful for document analysis. The basic idea of these approaches to modeling document content is that the probability distribution over words in a document can be expressed as a mixture model of K topics, where each topic is a probability distribution over words. In this paper, we use PLSA as the first topic modeling approach. In PLSA, an unobserved topic variable $z_k \in \{z_1, ..., z_K\}$ is associated with the occurrence of a word w_i in a particular document d_j. By summing out the latent variable z, the joint probability of an observed pair (d, w) is defined as

$$P(w_i, d_j) = P(d_j) \sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j), \quad (1)$$

where $P(w_i|z_k)$ is the probability of word w_i according to the topic model z_k, and $P(z_k|d_j)$ is the probability of topic z_k for document d_j. Following the likelihood principle, these parameters can be determined by maximizing the log-likelihood of a collection C as follows:

$$\mathcal{L}(C) = \sum_i \sum_j N_{ij} \log \sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j), \quad (2)$$

where N_{ij} denotes the occurrences of word w_i in d_j. The model parameters $\{P(w_i|z_k)\}$ and $\{P(z_k|d_j)\}$ are estimated by using a standard Expectation-maximization (EM) algorithm [10]. The estimated conditional probability (e.g., $P(z_k|d_j)$) is used to infer the cluster label for each document.

We use $\{P(z_k|d_j)\}$ as the weights of topics for document d_j, and the hidden topics can be regarded as clusters.

B. Biased Propagation based Topic Modeling

In the meanwhile, in order to emphasis more on event-related entities, we apply an entity-driven topic modeling approach described in our recent work [11], which is more suitable for the event extraction task because each event is associated with a set of entity arguments. For each document and its associated background metadata, we extract the named entities, such as persons and organizations, which may not only be highly correlated with the events but also cover the authors, publication venues and geographical places information of the documents.

We use a state-of-the-art bi-lingual entity extraction system [17] as our baseline to identify entities from English and Chinese. The system is trained on several years of ACE corpora, and can identify entities and classify them as persons, organizations, geo-political entities, locations, facilities, weapons and vehicles. For Chinese data, we applied the Tsinghua word segmenter [28] for pre-processing. The entity extraction system consists of a Hidden Markov Model (HMM) tagger augmented with a set of post-processing rules. The HMM tagger generally follows the Nymble model [3].

In general, the interactions among multi-typed entities play a key role at disclosing the rich semantics of the documents, and it is reasonable to build a 'virtual document' for each entity (e.g., person and organization) by aggregating their associated documents. Then, we obtain the term-person matrix U and the term-organization matrix V. In this way, documents and their associated entities are composed of words, so each of them can be decomposed by topic models, such as PLSA [16], respectively.

However, this method only considers the textual information while ignores the background network structures between documents and multi-typed entities. Here we apply the topic model with biased propagation (TMBP) [11] between documents and multi-typed entities to directly incorporate the heterogeneous information network with topic modeling in a unified way. The underlying intuition is that multi-typed entities should be treated differently along with their inherent textual information and the rich semantics of the relationships. For example, the topic distribution of an entity without explicit text information (e.g., person u_l) depends on the topic distribution of the documents that mention u_l. On the other hand, the topic of a document d_j is also correlated with its mentioned entities to some extend, but,

 ISBN: 978-1-61208-162-5

most importantly, its topic should be principally determined by its inherent content of the text. Thus, we define a regularization term as

$$R_U = \frac{1}{2}\sum_{i=1}^{|\mathcal{D}|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_l\in\mathcal{U}_{d_i}}\frac{P(z_k|u_l)}{|U_{d_i}|}\right)^2 + \frac{\tau}{2}\sum_{l=1}^{|\mathcal{U}|}\sum_{k=1}^{K}\left(P(z_k|u_l) - \sum_{d_i\in\mathcal{D}_{u_l}}\frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|}\right)^2. \quad (3)$$

A natural explanation of minimizing R_U is that entities should have similar topic distribution with their articles, and vice versa. Note that τ is the biased parameter. When $\tau \to \infty$, minimizing R_U will ensure the hypothesis that objects without explicit textual information are completely dependent on the estimated topic distributions of connected documents. Then the objective function R_U can be rewritten as

$$R_U = \frac{1}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_l\in U_{d_i}}\frac{P(z_k|u_l)}{|U_{d_i}|}\right)^2 \quad (4)$$

$$s.t. \quad P(z_k|u_l) - \sum_{d_i\in\mathcal{D}_{u_l}}\frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|} = 0. \quad (5)$$

Similarly, we could obtain the regularization term for other entities, e.g., organization v.

To incorporate both the textual information and the relationships between documents and multi-typed entities, we define a biased regularization framework by adding the regularization terms to the log-likelihood along with their constraints:

$$\mathcal{L} = \sum_i\sum_j N_{ij}\log\sum_{k=1}^{K}P(w_i|z_k)P(z_k|d_j) - \frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_j\in\mathcal{U}_{d_i}}\frac{P(z_k|u_j)}{|\mathcal{U}_{d_i}|}\right)^2 - \frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{v_j\in\mathcal{V}_{d_i}}\frac{P(z_k|v_j)}{|\mathcal{V}_{d_i}|}\right)^2 \quad (6)$$

$$s.t. \quad P(z_k|u_l) - \sum_{d_i\in\mathcal{D}_{u_l}}\frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|} = 0, \quad (7)$$

$$P(z_k|v_m) - \sum_{d_i\in\mathcal{D}_{v_m}}\frac{P(z_k|d_i)}{|\mathcal{D}_{v_m}|} = 0. \quad (8)$$

where λ is the regularization parameter which is used to control the balance between the data likelihood and the smoothness of topic distributions. We empirically set λ to 1000, and use generalized EM [6] for model fitting.

C. Performance Comparison

This new TMBP framework has proven much more effective than PLSA on scientific paper (DBLP and NSF award) domain. We believe that it is important to verify its effectiveness on the more general news domain before we apply it to enhance event extraction.

	NMI(%)	Accuracy(%)
PLSA	85.77	72.01
TMBP	**90.30**	**84.80**

Table I
TOPIC MODELING PERFORMANCE

We evaluate the topic modeling approaches on the Topic Detection and Tracking (TDT5) English corpus, which consists of data collected during April to September 2003, and taken from 7 sources, including Agence France Press, Associated Press, CNN, LA Times, New York Times, Ummah and Xinhua. It consists of 10,002 on-topic documents which are classified into 250 semantic topics. In our experiment, those documents appearing in two or more categories were removed, and only the largest 30 categories were kept, thus leaving us with 4,966 documents in total. There are 2,597 unique person entities, 2,161 unique organization entities and 1,199 unique geo-political entities embedded in these documents. There are in total 103,201 links among these entities and documents.

We adopted the following two standard scoring metrics, accuracy (AC) and normalized mutual information (NMI) [30] to measure the topic clustering performance:

$$AC = \frac{\sum_{i=1}^{n}\delta(a_i, map(l_i))}{n} \quad (9)$$

where n denotes the total number of objects; $\delta(x,y)$ equals 1 if $x = y$ otherwise 0; $map(l_i)$ is the mapping function [6] that maps each cluster label l_i to the equivalent label in the corpus.

Given two sets of document clusters C and C', the mutual information metric $MI(C,C')$ is defined as:

$$MI(C,C') = \sum_{c_i\in C, c'_j\in C'} p(c_i,c'_j)\cdot log_2\frac{p(c_i,c'_j)}{p(c_i\cdot c'_j)} \quad (10)$$

where $p(c_i)$ and $p(c'_j)$ denote the probabilities that a document arbitrarily selected from c_i and c'_j, and $p(c_i,c'_j)$ denotes the joint probability that a arbitrarily selected belongs to both c_i and c'_j at the same time. Let $H(C)$ denote the entropy of C, we use the normalized mutual information NMI as the $MI(C,C')$ normalized by $max(H(C),H(C'))$ which reaches from 0 to 1.

Table I shows the performance of PLSA and TMBP on document clustering for TDT5. We can clearly see that TMBP achieved much better performance than PLSA with both scoring metrics.

 ISBN: 978-1-61208-162-5

V. Applying Topic Modeling to Enhance Event Extraction

A. Data and Motivations

We use the 109 English newswire documents, 119 English web blogs and 238 Chinese newswire documents from ACE2005 training corpora to evaluate our approach. To simplify the analysis and experiment, we assign the most probable single topic cluster to each document. However, it is worth noting that although we discriminate the most relevant topic cluster and all other clusters, our documents are general news and blog articles and therefore each document may include more than one central topic. Therefore our method is not restricted to documents including single topics.

The most representative words in the resulting 5 (The value of 5 was arbitrarily chosen; variations in this number of clusters produce only small changes in performance) topics from our new topic model are presented in Table II and Table III, which coalesce around reasonable themes. For example, one can easily assemble possible topics correlating with certain types of events (e.g., "*attack*" events involving "*Israel*" in Cluster 1, "*meeting*" events involving "*Korean*" and "*nuclear weapons*" in Cluster 4, and "*Transaction*" and "*Justice*" events in Cluster 5).

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Palestinian	Iraq	Iraqi	north	court
Israel	war	forces	nuclear	dollars
police	United	Baghdad	Korea	year
Israeli	States	Iraq	weapons	appeal
people	Bush	troops	Korean	million
bank	Nations	city	talks	years
year	Iraqi	Saddam	officials	government
Monday	minister	military	Washington	convicted
killed	council	British	Putin	billion
west	resolution	American	south	sentence
security	security	officials	China	AFP
peace	country	regime	president	group
attack	president	army	United	Friday
city	role	Iraqis	Russian	April
university	Russia	Kurdish	States	life
officials	told	control	official	company
world	Tuesday	fighting	Pyongyang	case
attacks	France	northern	Russia	media
military	Washington	force	foreign	charges
house	government	Hussein	program	York

Table II
The Most Probable Words in 5 English Clusters

Across a heterogeneous document corpus, a particular verb can sometimes be an event trigger and sometimes not, and can represent different event types. However, within a cluster of topically-related documents, the distribution is much more focused. The word "*fire*" appears 81 times in the English training corpora and only 7% of them indicate "*End-Position*" events (a person stops working for an organization); while all of the "*fire*" in a topic cluster are "*End-Position*" events. The word "*Da*" appears 58 times in the Chinese training corpora and most of them indicate "*Attack*" events; while all of the "*Da*" in a topic cluster are "*Phone-write*" events (contact by phones or mails). Similarly, each entity tends to play the same argument role, or no role, for events with the same type in a topic cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Palestine	U.S.	company	team	China
Asia	president	court	game	Beijing
Israel	alliance	airline	Olympic	development
special	Bush	airplane	China	meeting
meeting	Relation	personnel	match	progress
conflict	State	three	world	national
Iraq	Europe	this year	coach	international
country	Germany	defendant	sports	country
army	problem	police	athletes	city
government	Yugoslavia	case	reporter	construction

Table III
The Most Probable Words in 5 Chinese Clusters (Translated)

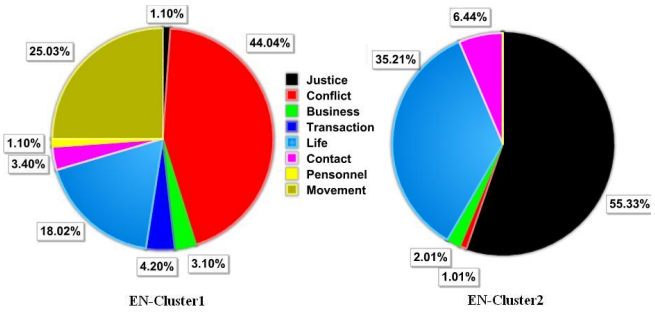


Figure 2. Event Distribution in Two English Clusters

Figure 2 and Figure 3 present the distributions of various event types in different topically-related document clusters. Although both EN-Cluster1 and EN-Cluster2 include certain amount of "*Life*" events (mostly "*Die*" subtypes), we can see that such "*Die*" events in EN-Cluster1 may have been caused by many "*Conflict*" events (44%), while EN-Cluster2 includes very few "*Conflict*" events. Instead, EN-Cluster2 includes many more "*Justice*" events (55%) than EN-Cluster 1(1.1%). Similarly, CH-Cluster1 includes a lot more "*Conflict*" events and fewer "*Movement*" events than CH-Cluster2.

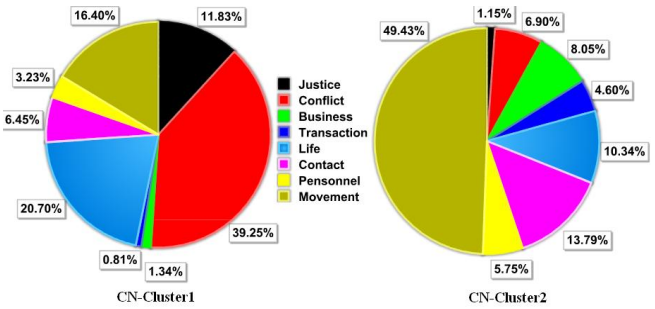


Figure 3. Event Distribution in Two Chinese Clusters

B. Topically Related Data is Better Data: Training Data Selection

Based on the intuition that the likelihood of a candidate word being an event trigger or an entity mention being an event argument in the test document is closer to its distribution in the collection of topically related documents than the uniform training corpora, we design the following active learning approach at selecting training data.

1. Apply topic modeling to the merged set of test documents and training documents to form various topic clusters.

2. For each test document d_i^j, which also denotes that d_i^j is the j^{th} document in the i^{th} topic cluster, the procedure can be formalized as follows.

(1) Add all topically-related training documents $\{d_i^k\}$ for training.

(2) Add all topically-unrelated training documents $\{d_l^m | l \neq i\}$ for training.

C. Cross-document Inference

We, generally, follow the hypotheses of "One Trigger Sense Per Cluster" and "One Argument Role Per Cluster" proposed by [18] to conduct cross-document inferences within each topic cluster. If we can determine the event type of a word or the role of an entity within a cluster of topically-related documents, this will allow us to infer its label in the test document. This method can fix event annotation errors produced by single-document extraction. Within each cluster, we conduct two types of inferences to favor interpretation consistency across documents:

- to remove triggers and arguments with low local and cluster-wide confidence;
- to adjust trigger and argument labeling to achieve cluster-wide consistency.

Ji and Grishman [18] required a large external collection of documents which were presumably topically related with the test set. In contrast, we found that the quality of extraction can be improved by partitioning the test set itself using topic models. In addition, they sent the candidate events produced by their baseline system as a query to an IR system to obtain the cluster for each test document. Therefore their inference performance may be limited by the quality of baseline extraction. In our topic modeling approach, we are able to take into account both candidate events and informative context words.

VI. Experimental Results

In this section, we present the results of applying this new topic modeling method to improve event extraction through extensive experiments.

A. Training Data Selection Results

For the active learning experiments, we setup a baseline passive learning approach by randomly selecting the same number of training documents for each test document.

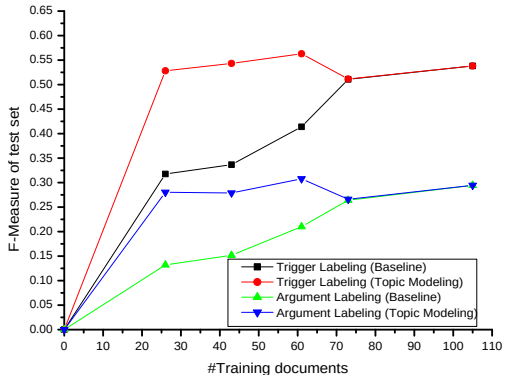


Figure 4. English Newswire Event Extraction

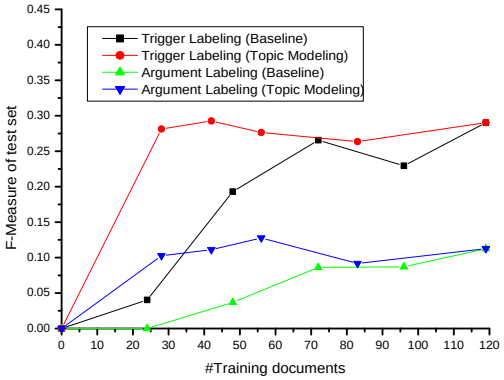


Figure 5. English Web Blog Event Extraction

Because of the data scarcity, leave-one-document-out cross-validation was used to train and test the event extraction systems. Figure 4, Figure 5 and Figure 6 present the F-measure results for both trigger labeling and argument labeling in two languages. The *x* axis in each figure shows the average number of training documents. The first point on each topic modeling curve indicates using all of the topically-related documents at once.

As expected, the baseline approach based on random

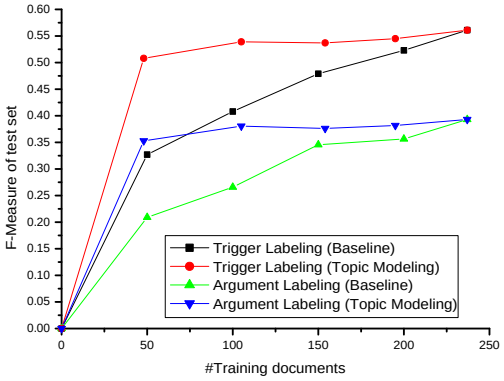


Figure 6. Chinese Newswire Event Extraction

System			Trigger Labeling			Argument Labeling		
			P	R	F	P	R	F
English newswire	Baseline		74.1	49.6	59.4	50.4	28.7	36.6
	Cross-doc Inference	IR	66.5	67.4	66.9	60.8	32.2	42.1
		Topic Modeling	**73.3**	**66.3**	**69.6**	**59.4**	**36.5**	**45.2**
English web blog	Baseline		43.2	29.4	35.0	20.9	15.6	17.9
	Cross-doc Inference	IR	38.5	42.6	40.4	30.2	21.4	25.0
		Topic Modeling	**41.9**	**43.8**	**42.8**	**32.3**	**23.8**	**27.4**
Chinese newswire	Baseline		78.8	48.3	60.0	60.6	34.3	43.8
	Cross-doc Inference	IR	69.9	62.3	65.9	67.5	38.3	48.9
		Topic Modeling	**76.5**	**61.9**	**68.4**	**66.4**	**42.4**	**51.8**

Table IV
CROSS-DOCUMENT INFERENCE RESULTS

selection produced almost linear increase as we add more and more training documents. In contrast, using only the topically-related documents, we can achieve comparable results as using the whole data sets. This indicates that our topic modeling based approach can dramatically speed up training data selection. Using the same amount of training data at the first point of each curve, topic modeling-based selection performs much better than random selection (18.1%-24.3% higher F-measure on trigger labeling and 10.3%-14.8% higher F-measure on argument labeling). For example, the named entity "*Putin*" appeared as different roles in various types of events in the English newswire data set, including "*meeting/entity*", "*movement/person*", "*transaction/recipient*" and "*election/person*". But the topic model was able to successfully divide them into different clusters, for example, "*Putin*" only played as an "*election/person*" in one cluster. In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on a document basis. The results show that the improvement using topically-related data over random selection is significant at more than 99.9% confidence levels for both trigger labeling and argument labeling in any language and genre. In fact, comparing the results using three clusters and all five clusters, we can see that adding topically unrelated documents can hurt performance for English.

B. Cross-document Inference Results

In order to conduct a fair comparison, we duplicated the IR-based clustering approach described in [18], and selected a similar size of data set for blind test (55 documents) for each setting (English newswire, English webblog and Chinese newswire). We then use all other documents in ACE2005 training corpora to train baseline event extraction systems.

Table IV shows the overall Precision (P), Recall (R) and F-Measure (F) scores for the blind test sets. Cross-document inference within topic clusters provided significant improvement over the baselines in both trigger labeling and argument labeling. We can also see that topic modeling-based approach achieved further improvement over the IR-based clustering approach. We conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on a document basis. The results show that the improvement using topic modeling over IR-based clustering is significant at more than 96% confidence levels for both trigger labeling and argument labeling for any language and genre.

VII. CONCLUSIONS AND FUTURE WORK

Most previous event extraction methods did not explore semantic links across multiple documents and background knowledge. In this paper, we described a novel genre-independent and language-independent topic modeling approach which structurally integrates interconnected entities and events across many documents. The resulting topic models were then used to effectively select training data and conduct global inference for event extraction. We expect this new framework will be also beneficial for other information extraction tasks. In the future we will aim to measure and reduce the impact of topic modeling errors on event extraction. We are also interested in extending our method to jointly enhance cross-lingual topic modeling [33] and event extraction.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. NSF IIS-09-05215, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. EAGER grant No. 1144111 and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

 ISBN: 978-1-61208-162-5

REFERENCES

[1] E. Agichtein and S. Cucerzan. Predicting accuracy of extracting information from unstructured text collections. In *CIKM*, pages 413–420, 2005.

[2] J. Allan. Topic detection and tracking: Event-based information organization. In *Springer*, 2002.

[3] D. M. Bikel, S. Miller, R. M. Schwartz, and R. M. Weischedel. Nymble: a high-performance learning name-finder. In *ANLP*, pages 194–201, 1997.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. L. Boyd-Graber, D. M. Blei, and X. Zhu. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, pages 1024–1033, 2007.

[6] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM*, pages 911–920, 2008.

[7] Y. S. Chan and D. Roth. Exploiting background knowledge for relation extraction. In *COLING*, pages 152–160, 2010.

[8] Z. Chen and H. Ji. Can one language bootstrap the other: A case study on event extraction. In *HLT-NAACL 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, 2009.

[9] Z. Chen and H. Ji. Collaborative ranking: A case study on entity linking. In *EMNLP*, 2011.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38*, 1977.

[11] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *KDD*, pages 1271–1279, 2011.

[12] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *IJCAI*, pages 1034–1041, 2005.

[13] N. Fernandez, J. A. Fisteus, L. Sanchez, and E. Martin. Webtlab: A cooccurencebased approach to kbp 2010 entity-linking task. In *TAC 2010 Workshop*, 2010.

[14] P. Gupta and H. Ji. Predicting unknown time arguments based on cross-event propagation. In *ACL/AFNLP (Short Papers)*, pages 369–372, 2009.

[15] V. Ha-Thuc, Y. Mejova, C. Harris, and P. Srinivasan. A relevance-based topic model for news event tracking. In *SIGIR*, pages 764–765, 2009.

[16] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[17] H. Ji and R. Grishman. Improving name tagging by reference resolution and relation detection. In *ACL*, 2005.

[18] H. Ji and R. Grishman. Refining event extraction through cross-document inference. In *ACL*, pages 254–262, 2008.

[19] R. Jones, R. Ghani, T. Mitchell, and E. Riloff. Active learning for information extraction with multiple view feature sets. In *ECML Workshop on Adaptive Text Extraction and Mining*, 2003.

[20] LDC. Automatic content extraction (ace): http://www.itl.nist.gov/iad/mig/tests/ace/. 2005.

[21] S. Liao and R. Grishman. Using document level cross-event inference to improve event extraction. In *ACL*, pages 789–797, 2010.

[22] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR*, pages 186–193, 2004.

[23] S. Patwardhan and E. Riloff. Effective information extraction with semantic affinity patterns and relevant regions. In *EMNLP-CoNLL*, pages 717–727, 2007.

[24] S. Patwardhan and E. Riloff. A unified model of phrasal and sentential evidence for information extraction. In *EMNLP*, pages 151–160, 2009.

[25] A. Rahman and V. Ng. Coreference resolution with world knowledge. In *ACL*, pages 814–824, 2011.

[26] B. L. Smith and J. T. MacGregor. What is collaborative learning? In *Collaborative Learning: A Sourcebook for Higher Education*, 1992.

[27] J. Sun, T. Wang, L. Li, and X. Wu. Person name disambiguation based on topic model. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2010.

[28] M. Wan and Z. Luo. Study on topic segmentation method in automatic abstracting system. In *Natural Language Processing and Knowledge Engineering*, 2003.

[29] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[30] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.

[31] R. Yangarber. Verification of facts across document boundaries. In *International Workshop on Intelligent Information Access*, 2006.

[32] J. M. Zacks and B. Tversky. Event structure in perception and conception. In *Psychological Bulletin, Vol. 127 (2001), pp. 3-21*, 2001.

[33] D. Zhang, Q. Mei, and C. Zhai. Cross-lingual latent topic extraction. In *ACL*, pages 1128–1137, 2010.

[34] S. Zhou, K. Li, and Y. Liu. Text categorization based on topic model. In *Lecture Notes in Computer Science, Volume 5009/2008, 572-579*, 2008.

 ISBN: 978-1-61208-162-5

Mining Ice Hockey: Continuous Data Flow Analysis

Adam Hipp
Applied Computational Intelligence Laboratory
University of Cincinnati
Cincinnati, Ohio 45220
hipp.adam@uc.edu

Lawrence J. Mazlack
Applied Computational Intelligence Laboratory
University of Cincinnati
Cincinnati, Ohio 45220
mazlack@uc.edu

***Abstract*—Ice hockey is relatively under computationally analyzed. Possibly this is because ice hockey is a continuous flow game with relatively few major events (goal scoring) while most of the other games that have been data mined can be described as being a series of clearly bounded events. This work describes needs of data mining ice hockey statistics to quantify the contribution of individual hockey players to team success. Large databases of ice hockey statistics for the collegiate and professional levels can be accessed to perform this work. The goal is to use ice hockey statistics and computational methods to help make personnel decisions at both the coaching and franchise management levels. This type of work has the potential to encourage new avenues of sports statistics research, as well as statistical research and data mining, in general.**

Keywords-data mining; continuous flow analysis; ice hockey

I. Objectives and Overview

Data mining techniques have been developed to take large bodies of statistics and reduce them to interesting patterns and relationships using computers much faster than a human could on his own. Data mining techniques have been successfully applied to many types of large databases, and an emerging application of data mining is sports statistics. Sports generally lend themselves well to data mining since most sports contain large amounts of statistics that have been kept over a long period of time.

The purpose of this work is to data mine ice hockey statistics to quantify the contribution of individual hockey players to team success (i.e., winning). Large databases of ice hockey statistics for the collegiate and professional levels can be accessed to perform this work. The long term goal is to use ice hockey statistics and computational methods to help make personnel decisions at both the coaching and franchise management levels. For example, applications of this work could be used to help a coach decide which players on his team should receive more playing time, or to help a team executive decide who on his team should be traded and for whom from another team.

The central hypothesis is that currently available ice hockey statistical databases, which contain statistics compiled for individual players, can be computationally used to aid in making effective decisions about ice hockey players. The rationale behind this hypothesis is that computational approaches to player statistics in other sports in a similar fashion have been successful in the past, and as stated, there is a large amount of hockey statistics available for analysis. There are two specific aims:

- **Create a computer model that takes ice hockey statistics as an input and scores each player's contribution to their team**. The working hypothesis is that currently available ice hockey statistics can be used to effectively compare players to one another on a single, useful, and unbiased scale.
- **Create a more robust computer model that takes multiple players' statistics into account to quantify how effective multiple players are together.** The working hypothesis is that currently available ice hockey statistics can be used to score a "squad" contribution level on a single, useful, and unbiased scale.

While data mining methods have been applied to some sports with success, very little research has been done specifically on ice hockey. The results add to the emerging field of sports data mining.

II. Significance

Data mining and statistical analysis has been applied to other major sports with some success, but very little research has been done with ice hockey statistics. The argument could be made that this is because ice hockey does not lend itself to this type of analysis as easily as sports such as basketball, baseball, football, and soccer. The difference between ice hockey and these other games can be identified by the internal game flow. In baseball and football, action is broken into a sequence of separate plays; the result of what each player does in a particular play is easily quantified, for example: bases accumulated on a hit in baseball or yards gained on a single carry by a running back in football. In basketball, plays are not as broken up, but scoring is frequent, causing more statistics to be accumulated for each player throughout the course of the game. Hockey is a low scoring, relatively constant flow game.

There are a significant amount of hockey statistical categories, even if the number of categories is less than the other games. The primary statistics a player accumulates are goals, assists, points (goals + assists), and +/- (difference between team goals scored and team goals allowed while the player is on the ice). Other lesser known statistics have been compiled as well. The website of the National Hockey League (NHL), NHL.com, keeps detailed player statistics dating back to 1997, and appends these basic statistical categories with penalty minutes, power play goals, shorthanded goals, game winning goals, game tying goals, overtime goals, shots, shooting percentage, time on ice, shifts per game, and face-off win percentage. There is

 ISBN: 978-1-61208-162-5

potential in the accumulated available data to quantify player contributions using more complexity than just goals and assists. In depth statistical analysis of a given set of sports statistics can lead to the creation of new statistical categories, causing useful analysis to build upon itself.

Based on the successful application of statistical techniques and data mining to other sports and the untapped potential of ice hockey statistics, the significance of the gap in knowledge of applying these techniques to ice hockey is made clearer.

This work is expected to improve the decision making abilities of coaches, team executives, and other people involved in management of an ice hockey team. There are many types of benefits that can come out of the new knowledge generated by this work.

One group of people who will benefit from the work is ice hockey coaches. The largest beneficiaries would be hockey coaches at the highest level, professional or international team coaches. This is because the most statistics and interest are available at the professional level. The results of this work can be used to help a coach decide which players should receive more playing time, or when different players can be ideally used in different situations. The results will also help a coach identify players who are undervalued on his roster or what players would benefit most from what type of development. Over time, a coach may learn to identify traits in a player that may have been suggested by the statistical analysis, which can improve a coach's talent and instincts.

A second group of people who will benefit from this work are team executives If this analysis identifies a player on another team who is undervalued by his team, a team executive may use this information to propose a trade for the undervalued player without having to give up too much for a potentially good player. This work can also help team executives decide what players to draft from colleges and other lower leagues without having to rely purely on scouts who can only investigate a limited number of players. This work can help teams discover players who might otherwise not have been given a chance.

Another benefit of this work is the benefit to the followers of ice hockey. By having better players and better teams taking the ice every day, the quality of the game will be improved, which will make the game more enjoyable for the fans. Beyond this, many fans of sports enjoy hearing statistical analyses of the games they follow, and this work will open the doors to new types of hockey analysis. More interesting statistics will help make ice hockey even more interesting and generate more avid followers of the game.

III. BACKGROUND

There has been some research into applications of statistics and data mining to sports. This research has approached the problem of modeling sports situations and performance in a variety of ways. One vein of sports research that can be found in the literature is research which attempts to statistically model in-game situations. This can be most commonly seen with soccer, since, like hockey, soccer is a constantly moving, low scoring game, leading to less statistics to be generated.

For example, Wang and Wang [1] used the Apriori algorithm [2] to determine association rules between different types of technical movements in soccer. The idea behind this research was to determine what the most common chains of ball movements that occur during a particular game are. This type of research can be applied to the ball movement patterns of a particular team to determine what kind of defense to play against the team, or to help players predict what the opposing team's next move may be after something occurs based on common occurrence chains determined from application of this research. The only thing required would be for someone to watch the opposing team play and keep a record of when each technical ball movement occurred throughout the game, generating a long chain that represented the entire game for the team. By watching many games, a database of ball movement can be created and then mined using the research as presented.

A similar approach was taken by Chai [3] to mine patterns from data of soccer using serial data, pass time, and ball possession time. Serial data refers to the pattern of which player possesses the ball in a given chain of possession for a team. Chai applied data mining techniques to the serial data to determine what chains of possession between players happened often, incorporating both teams with specific players into each pattern. This work can be used to show what trends of ball possession happen when a specific team is in control. Also, it can help to identify weaknesses within that team's strategy by incorporating the players on the opposing team. If a particular chain of possession for a team was often broken up by a particular player on the opposing team, a future opponent can use that information to adjust its strategy.

Beyond just the actual patterns of player possession determined, the pass time and ball possession time was used to create a time series model of a particular game, which led to other association rules about each interesting chain of player possession which gave some insight into why events occurred. For example, a chain of four specific players may end in a steal by the opposing team often, but the reason why would be up to speculation from that information alone. However, by knowing that the specific chain of players led to longer pass times, it can be inferred that these players tend to make long, dangerous passes to one another, which can be used to help strategize against that team in the future. [3]

Nunes and Sousa [4] applied different data mining and data visualization techniques to available data relating to the European championship soccer matches over the history of the tournaments. Using in game data such as cards given, goals scored, and substitutions, these techniques were able to confirm the types of trends and patterns one may expect for a soccer match. For example, the research showed that substitutions begin occurring towards the end of the first half, with very few substitutions early in the match. Since when a player is substituted out in a soccer match, he is done for the rest of the game, it makes sense that substituting early in the match would be uncommon.

 ISBN: 978-1-61208-162-5

Halftime is often a good time to make a substitution since it gives coaches and players time to refocus their strategy with the break.

Nunes and Sousa [4] applied the Apriori algorithm to match specific information from the tournament's history in the hopes of finding interesting and unexpected association rules, but the results of this portion of the research did not yield interesting results. Instead, trivial association rules were found. The research went on to apply classification data mining techniques to matches in different countries but it was again unable to find many useful classifications in terms of predicting the outcome of soccer matches. Data visualization proved to be the most useful of the statistical techniques applied to the data in this research.

There have been various levels of success applying game-model statistical research to soccer matches. Soccer and ice hockey share many traits in terms of how the game can be modeled, and one avenue of ice hockey statistical research that may be worth pursuing is the one taken for soccer research shown here. However, that is not the type of data mining for ice hockey that is described here. Data mining that makes use of a greater number player and team statistics in a more generalized fashion, which is more similar to the work described here, can be found in the literature.

Maheswari and Rajaram [5] decided to apply data mining techniques to cricket data, noting that cricket lends itself to large accumulations of data which cannot be completely analyzed by a human in a reasonable amount of time. The approach in this work straddles the line between game flow modeling, similar to the soccer research just presented, and purely accumulated statistical modeling. In cricket, one team bowls the ball towards the batter, who is a member from the other team. The batter is on offense and the bowler is on defense. Play can be broken up by each bowl, where the bowler bowls the ball and the hitter hits it. A sequence of events will occur which may or may not result in runs being scored, and then the process repeats with a new bowl. Since cricket is a game where many short plays occur, similar to baseball, association between statistical events is much clearer to infer. This work uses data mining algorithms to find association rules between different statistical categories in cricket for different players. This type of information can be used to find a player's strengths and weaknesses to different approaches from the opposition. For example, the type of bowl to which a hitter is weak could be discovered through an application of this work helping the defensive team strategize as to how to bowl to that particular player.

Recently, Wang, Jie, and Zeng published a paper related to their work on development of an interactive baseball and softball statistical gathering and analysis computer program [6]. Fast and Jensen [7] applied data mining techniques to broad NFL team statistics, such as wins, losses, points scored and points allowed, as well as the history of NFL coaching performances and associations between coaches. Coaching associations comes in the form of who was an assistant under whom, to give an idea of who learned from successful coaches and went on to become successful himself. Based on statistical team success and coaching associations, the research was able to show that the history of the coaching staff for a given team had a strong association to whether that team made the NFL playoffs in the given year. This type of research could be used to help identify up and coming or undervalued coaches quickly, so a franchise or school would move in to acquire this coach before other teams or schools start to take notice. Conversely, this research could also be used to show what coaches may be given more credit than they deserve.

Romer [8] applied statistical analysis and dynamic programming to model NFL football game situations with the goal of determining optimal strategy on fourth-down plays. The generally accepted strategy in football is for the team with the ball to punt in many given situations when the team has the ball on fourth down. Romer's analysis shows that conventional knowledge may be approaching the game too conservatively, meaning that teams should be more willing to "go for it" on fourth down to try to keep possession of the ball if they are able to gain enough yards.

There has even been sports research from a more media-driven point of view, such as the research by Dao and Babaguchi [9]. This research uses a temporal representation of small events during the course of a sporting event to predict what type of event will occur next. The results of the research could be useful for media coverage of sporting events, to help a system automatically identify when a replay was likely to be displayed to viewers, or for the system to understand what kind of event is happening on the field based on the positioning of cameras at the current time and moments leading up to the current time. This type of automatic detection could be used to improve the speed and quality of media coverage of sporting events. Since this type of detection would be nearly impossible to have perfect results, the best application would likely to let the system automatically suggest to an interactive user what the next step in coverage to take should be, which could speed up the decision making process within the media and thus enhance the viewer experience.

An application of data mining takes a similar approach to ours was done by Smith, Lipscomb, and Simkins [10]. In their work, the goal was to use data mining of individual baseball pitcher statistics to predict the winner of the annual Cy Young award. The Cy Young award is given to the pitched voted to be the best pitcher in each of the two leagues that make up Major League Baseball. The application in this research chose the top 10 starting pitchers in terms of wins and top 10 relief pitchers in terms of saves as candidates for the Cy Young in the given season. The individual statistics available for each candidate are then run through a data mining algorithm that is a Bayesian classifier. Applying the classifier to data from 1967 through 2006, the correct pick of Cy Young winner from each league happened frequently.

In the 1990's, IBM developed a program called Advanced Scout [11] that was created to "seek out and discover interesting patterns in game data." Advanced Scout was a specific data mining application for NBA basketball, and was distributed to many NBA teams during the 1990's.

 ISBN: 978-1-61208-162-5

Using input data from a given game of basketball between two teams, which includes very specific input data such as "who took a shot, the type of shot, the outcome, any rebounds, etc." The data is also temporal, showing when each data point occurred during the game. The IBM software is then able to apply data mining algorithms to this data and find interesting and useful patterns and associations. Since the data is temporal, video can accompany a game analysis, so when a particular pattern is shown to occur at a time point, a coach can then pull up the video to actually see what the statistics are saying. This type of interaction can be very valuable to a coach, helping him to learn to spot important patterns and behaviors on his own, as well as to better understand those identified by the data.

Some of the general data mining queries involve "either field goal shooting percentage to detect patterns related to shooting performance, or possession analysis to determine optimal lineup combinations." Advanced Scout takes an approach similar to the cricket application by Maheswari and Rajaram [5], but with more detailed input data and more rigorous analysis.

Another broader-focus type of sports data analysis is known as "sabermetrics." The beginnings of sabermetrics are well explained by Schumaker [12]: "In 1977, Bill James began publishing his annual *Bill James Baseball Abstracts*. These abstracts were used as his personal forum to question many of the traditional baseball performance metrics… James continued to publish his annual compendium of insights, unorthodox ranking formulae and new statistical performance measures which he called *sabermetrics*." Baseball is a statistics rich game, and it is now accepted that rigorous statistical analysis is necessary to remain competitive at the professional franchise level of baseball. However, James' sabermetrics were not even strongly considered by decision makers in Major League Baseball until 2002, when the Oakland Athletics began to incorporate some of James' ideas. The A's used some of James' performance measurements to help decide whom to draft, and the strategy marked a turnaround for the A's franchise. The A's general manager, Billy Beane, "discovered that by carefully selecting players in the draft, the A's could lock-in players that were oftentimes overlooked by other clubs into long contracts that paid little money and thus develop this into a strategy to compete with larger payroll teams. It was simply a matter of picking the right players, which sabermetrics could make easier."

The traditional baseball statistics are categories like runs batted in (RBI), hits, singles, doubles, triples, home runs, and stolen bases for field players, and categories like earned run average for pitchers, which takes the number of earned runs a pitcher gives up and normalizes it by 9 innings to give a measure of the average number of runs that pitch would give up in a normal nine inning game. James came up with different categories that helped to better quantify player performance, such as On-Base Percentage, which determines what percentage of plate appearances for a batter results in the batter reaching base, and strikeout to walk ratio for pitchers.

Schumaker [12] explained how Dean Oliver took a similar approach to basketball statistics as James did to baseball beginning in the 1980's. Oliver took more of a team-based statistical approach, but still was able to make contributions to analysis of basketball by helping to identify player contribution and even measure how well players worked with one another on the court.

As the success of analysis such as that of James and Oliver began to be noticed, a sudden revolution in sports data analysis began and is continuing today. A sabermetrics type of analysis is beginning to be applied to football as well, which has the benefit of many types of statistics easily able to be accrued due to the stop-and-go nature of the game, but the drawback in that there are only 16 games in an NFL season, while there are 162 games in a baseball season and 82 games in an NBA season. As a comparison, there are 82 games in an NHL ice hockey season as well. [12]

There have been plenty of statistical computer tools that have been created to access and analyze the large amount of sports statistics available. Schumaker [12] provides a short summary of different programs, some of which are for a specific sport, while others can be used for many different games. One example is Digital Scout, which is an adaptable piece of software that can be used for record keeping and creating custom reports, "such as baseball hit charts, basketball shots, and football formation strengths." Another is SportsVis, which creates graphical representations of sports data quickly to help people uncover trends or problems.

Schumaker [12] delved into research tools and methods created for a large variety of sports, but ice hockey information is noticeably sparse. Comprehensive statistical data about players and teams in hockey can be found in two major online resources are supplied by the NHL [13] as well as an outside, independent resource [14]. Interactive graphics that can be used to identify where events occur on the ice and when can be found are also supplied by the NFL [13]. Similar to basketball, a shot-taking map can be drawn on a graphical representation of the ice/court, and the visual data can infer trends or patterns about players or a team [15].

IV. DATA

There are two specific objectives of the work. The first objective is to create a computer model that takes ice hockey statistics as an input and scores each player's contribution to his team. The second objective is to create a robust computer model which takes multiple players' statistics into account to quantify how effective multiple players are together.

To accomplish both of these objectives, a large amount of data needs to be accessed. NHL.com has comprehensive statistics reaching back to the 1997-1998 season for every player, including Games Played, Goals, Assists, Points, +/-, Penalty Minutes, Power Play Goals, Short Handed Goals, Game Winning Goals, Overtime Goals, Shots, Shooting Percentage, Time on Ice per Game, Shifts per Game, and

 ISBN: 978-1-61208-162-5

Face-off winning percentage. The website also has game-by-game statistics and team statistics.

Since the first goal is to score a player's contribution to team success, both the individual player statistics and team statistics will be important to consider. For example, if one player has scored many more goals then another, but the high-scoring player is on a team with many more losses than the lower-scoring player, that doesn't necessarily mean that the higher-scoring player is a better contributor to team winning. There could be something else the player does that makes it more difficult for his team to win, which may even be related to his higher scoring. If the higher-scoring player has a tendency to play too aggressively, he will score more goals, but he'll also put his teammates in more shorthanded situations, either through penalties or by trying to take the puck too far on his own, causing him to turn the puck over more often and lead to breakaways for the opposing team. The basic idea is to use player statistics and the team statistics accumulated at the same time. With a large database to work with, the hope is to mine interesting association rules. However, the approach for mining these rules will not be straightforward. In fact, players may need to be grouped differently depending on their position. For example, penalty minutes accrued by a center may have an association with accumulating losses for a team, but penalty minutes accrued by a defender may have an association with accumulating wins for a team. This is a possibility due to the different nature of the positions. A center's job is to win face-offs and score goals; a defender's job is to stop the opposing team from scoring. Generally, a rougher, more physical player is a desirable attribute in a defender, and those types of players will tend to accumulate more penalty minutes. Again, this is speculation, and quite the opposite may result from the analysis.

Once a variety of data segmentation and approaches to mining association rules between player statistics and team wins have been run, the intention would be for a comprehensive formula that scores each player's contribution to team success based on their statistics. The formula will be regardless of who the player is or what team the player is on. This way, the value of two players from very different backgrounds, including position differences and team history, can be compared directly in an unbiased way. If an effective formula can be derived, it could be used to identify a player who makes strong contributions to his team but has been stuck on unsuccessful teams due to the players around him, which could be used to the advantage of both the team who has the player and a different team who would want to acquire him.

To accomplish the second objective, a similar approach is taken as with the first, using pairs of defenders and lines of forwards, where a line consists of two wings and a center. The groups of players to be used for creating the association rules will be chosen by researching commonly used lines on teams throughout the time period covered by available data. This is because randomly choosing players to create a "line" for each team would work statistically, but could cause misleading results since the players may not necessarily work together. The goal is to derive association rules that may have a form like "If defender 1 and defender 2 both have an average of more than 2 penalty minutes per game with an average time on ice of at least 20 minutes per game, then wins occur."

Similar to the first process, if a useful set of interesting association rules can be derived, the next step would be to develop a mathematical formula using the statistics of a pair or set of three players and "score" the group of players. If actual scoring cannot be accomplished, at least a set of player types that create an idea pair of defenders or line of forward could be suggested with statistical support. With the complex level of analysis required for combining players, even suggesting what "groups of statistics" for each of two or three players on a line could be very valuable in constructing lineups for a game or figuring out who to draft or acquire for a team, based on the current roster.

VI. CONCLUSION

There is sufficient precedent to show the success and benefits of sports statistics and data mining analysis. Ice hockey is relatively under computationally analyzed. Possibly this is because ice hockey is a continuous flow game with relatively few major events (goal scoring) while most of the other games that have been data mined can be described as being a series of clearly bounded events. This work described the needs of data mining ice hockey statistics to quantify the contribution of individual hockey players to team success. Large databases of ice hockey statistics for the collegiate and professional levels can be accessed to perform this work. The goal is to use ice hockey statistics and computational methods to help make personnel decisions at both the coaching and franchise management levels. Many people from the walks of academia and business, as well as hockey enthusiasts, will benefit from the results of this research. This type of work also has the potential to encourage new avenues of sports statistics research, as well as statistical research and data mining, in general. It is an important societal goal to find an effective way to draw more people into studying fields involving math and science.

REFERENCES

[1] B. Wang and L. Wang, "Research of Association Rules in Analyzing Technique of Football Match," presented at Second International Conference on Power Electronics and Intelligent Transportation Systems, 178-180, 2009.

[2] R. Agrawal and R. Srikant, "Fast Algorithms For Mining Association Rules," presented at 20th VLDB Conference, 487-499, 1995.

[3] B. Chai, "Time Series Data Mining Implemented on Football Match," *Applied Mechanics and Materials*, vol. 26-28, pp. 98-103, 2010.

[4] S. Nunes and M. Sousa, "Applying Data Mining Techniques to Football Data from European Championships," presented at Conferência de Metodologias de Investigação Científica (CoMIC'06), 4-16, 2006.

[5] P. Maheswari and M. Rajaram, "A Novel Approach for Mining Association Rules on Sports Data using Component

 ISBN: 978-1-61208-162-5

Analysis: For Cricket match perspective," presented at 2009 IEEE International Advance Computing Conference, 2009.

[6] G. Wang, S. Jie, and F. Zeng, "Design and Realization of Baseball and Softball Match Data Analysis Information System," *Advanced Materials Research*, vol. 187, pp. 353-357, 2011.

[7] A. Fast and D. Jensen, "The NFL Coaching Network: Analysis of the Social Network Among Professional Football Coaches," presented at AAAI Fall Symposia on Capturing and Using Patterns for Evidence Detection, 2006.

[8] D. Romer, "It's Fourth Down and What Does the Bellman Equation Say? A Dynamic-Programming Analysis of Football Strategy," 9024, 2003.

[9] M. Dao and N. Babaguchi, "Sports Event Detection using Temporal Patterns and Web-casting Text.," presented at First ACM Workshop On Analysis And Retrieval Of Events/Actions And Work Flows In Video Streams (AREA '08), 2008.

[10] L. Smith, B. Lipscomb, and A. Simkins, "Data Mining in Sports: Predicting Cy Young Winners," *Journal of Computing Sciences in Colleges*, vol. 22, pp. 115-121, 2007.

[11] I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pretap, and K. Ramanujam, "Advanced Scout: Data Mining And Knowledge Discovery In NBA Data," *Data Mining and Knowledge Discovery*, vol. 1, pp. 121-125, 1997.

[12] R. Schumaker, O. Solieman, and H. Chen, "Research in Sports Statistics," *Sports Data Mining*, pp. 33-50, 2011.

[13] http://www.nhl.com/

[14] http://www.hockey-reference.com/

[15] H. Chen, R. Schumaker, and O. Solieman, "Data Sources for Sports," *Sports Data Mining*, pp. 25-31, 2011.

 ISBN: 978-1-61208-162-5

Knowledge Management System IMMPETUS (KMSI) - Connoisseur

Osman Ishaque
Department of Automatic Control and Systems Engineering
The University of Sheffield
Sheffield, UK
E-mail: O.Ishaque@sheffield.ac.uk

Mahdi Mahfouf
Department of Automatic Control and Systems Engineering
The University of Sheffield
Sheffield, UK
E-mail: m.mahfouf@sheffield.ac.uk

George Panoutsos
Department of Automatic Control and Systems Engineering
The University of Sheffield,
Sheffield, UK
E-mail: g.panoutsos@sheffield.ac.uk

Lucian Tipi
Sheffield Business School
Sheffield Hallam University
Sheffield, UK
Email: L.Tipi@shu.ac.uk

***Abstract*— KMSI (Knowledge Management System IMMPETUS) is a web based portal which facilitates IMMPETUS (Institute for Microstructural and Mechanical Process Engineering: The University of Sheffield) users, potentially providing a web area where data mining, data archiving, data retrieving are systematically carried out. It is a versatile web portal which can handle raw data, research knowledge and sharing of information with transparent access by the whole of the IMMPETUS research team. This includes internally and externally generated data and knowledge, past and present. Hitherto, KMSI has focused on hierarchical structures with bi-directional communication between the various levels. KMSI's previous configuration had limited functionality in terms of collaboration among project groups and across the whole of the IMMPETUS. Contributions of knowledge by individual members are now promoted as this is vital for Information Management and help users to track their project work. In this paper, we explain how Microsoft SharePoint 2010 framework technology can easily be used to develop; a versatile web portal to systematically include optimal experimental design, data analyses, collaboration, and overall knowledge management.**

Keywords-information storage; data sharing; data mining, knowledge management; decision making.

I. INTRODUCTION

"IMMPETUS is a research institute providing its wide range of services to integrate metallurgical, mechanical and thermal considerations in developing soundly based models for process planning and control to achieve target microstructures and product properties within increasingly fine tolerances and greater efficiency" [1]. IMMPETUS constitutes itself with the help of three disciplines of control systems, mechanical systems and engineering materials to produce world leading research. IMMPETUS, as an institute, is internationally renowned for its innovative approach for the study of the thermomechanical processing of metals, which includes Steel, Aluminium and Titanium alloys. IMMPETUS research emphasizes on; (1) the deeper physical understanding of the thermo mechanical behaviour, and the development of the associated models with adequate accuracy and transparency, (2) the identification of optimal processing routes to achieve pre-defined microstructural and mechanical properties [1].

A. Requirements

Over a period of time spanning more than a decade IMMPETUS has since evolved and the number of research projects carried out under its umbrella has significantly increased. Academic supervisors are responsible for managing various research projects simultaneously from microstructure to modelling and control. During the project development phase collaboration and information sharing among the various groups is essential; once a project is completed all the artefacts and mature data are needed for future usages of development and research work.

The collection of data, analysing the results and sharing the right information across the research staff is a vital task for all researchers.

Our aim with this work was to develop, using the latest technologies, a system; (1) which easily scales across the IMMPETUS potentially giving any research project a dedicated web-based work area where one can store project artefacts and will be able to quickly retrieve the desired content whenever it is needed, (2) a team-oriented collaboration tool to help project members to monitor project progress, (3) potentially providing knowledge sharing, document storage and advanced user-controlled features, and (4) services to store and analyse experimental data.

 ISBN: 978-1-61208-162-5

II. KNOWLEDGE MANAGEMENT SYSTEM

Knowledge management systems (KMS) are IT systems developed to store, manage and share on organization's knowledge. These IT-based systems are significantly important to facilitate and enhance the processes of knowledge creation, storage and retrieval [2].

It is a challenging task to capture all the data from various sources and to disseminate the most adequate information using IT systems. Gathering all the data from various sources and processing it in order for it to have some meaningful information is in itself a thought-provoking job.

IT systems are becoming *the tool* relied upon by many organisations around the globe to decide on how to manage electronic records [3].

The list of vendors who provide various software solutions to manage data and information is rather long but difficult to report in full. Gartner has published a comparison of 'Enterprise Content Management Systems' [4] being provided by some of major vendors. It is however worth noting that it is not always straightforward to purchase ready-made systems that suit one organization's needs, but many customising work will be needed.

III. KMSI CONFIGURATION

KMSI's configuration is fundamentally organised so as to gather the experimental data, expert knowledge, preliminary results, mature data, and document storage. Because group collaboration is also promoted, it has the potential of providing a robust infrastructure for storage and collaboration, a foundation platform for a web-based project areas (workspaces) and services, and a more efficient way of using tools for information-sharing that help users stay connected and informed across the whole of IMMPETUS.

A. Why Dynamic Knowledge Management System ?

Dynamic Knowledge Management Systems are concerned with the sharing of knowledge that already exists but also focus on the production of new knowledge [5].

In IMMPETUS, every research project has its own boundaries and requirements; therefore the KMSI system architecture is designed to handle various types of information in a holistic way. A platform was needed where all past research information is located and current projects artefacts can be stored, content can be accurately classified, and information sharing can be made easy in a fashion where group collaboration is a core element. When new research projects are initiated in IMMPETUS more web-based project areas are required within existing KMSI structure having all modules which are vital to complete that specific project

B. Putablity and Findability

The information process should be made easy for uploading the content and retrieving the desired information effectively which represents hub of knowledge management.

"*Putability is the quality of putting content into an information management and retrieval system with the correct metadata."* [6].

"*Findability is the quality of being locatable or navigable*" [6], in other words it is the process which helps to find desired content easily even intuitively.

It is vital for better decision making to obtain the desired information whenever it is needed. KMSI is designed to put and locate the desired content quickly using 'SharePoint 2010 Search' [7] to best fit with modern age IT requirements.

C. Technology Used

Microsoft latest software and technologies are used to build such a robust portal. Some of the details about the hardware and the software used to configure KMSI are as follows:

a) Hardware

- 2 Processors, 2.53.Ghz, Core-2 Quad
- Ram 16 GB
- 2 Hard disks 250GB each for Operating System, configured on Raid-1
- 3 Hard disks 500GB each for data base, configured on Raid-5

b) Software

- Microsoft Windows Server 2008 R2 (64 bit)
- Microsoft SharePoint 2010 (64 bit)
- Microsoft SQL Server 2008 (64 bit)

D. KMSI's Architecture

KMSI's architecture has been designed so as to accomplish all IMMPETUS information management requirements; its dynamic structure can easily be extended for developing future web-based project and knowledge sharing areas. The contribution of knowledge by individual members has also been added as it is vital for Information Management and storage. Latest technologies are used to enhance the usability and boost the ability of storing, sharing and publishing information. The Server Administrator can create as many areas as required which inherit features from the core platform. The baseline hierarchy aims at providing more flexibility and knowledge sharing between members using more intuitive ways. Information retrieval from experimental equipment is integrated in a customizable manner which has the flexibility to alter interfaces as per diverse project requirements. Architecture has the potential to import experimental results directly from some modern equipment using their API (application programmable interface) which enhances its productivity.

The three main different web applications have been created on IIS (Internet Information Services) web server having their own separate data-bases to keep the process isolation. The web applications within KMSI are as:

(1) Data Libraries Area: A publication web site which archives all important information using modern web interface. Information stored here is available to all IMMPETUS members.

(2) Projects Area: The web application designed within KMSI, having multiple project sites and sub-sites. Each IMMPETUS core project has its own workspace or web site.

 ISBN: 978-1-61208-162-5

Content modification rights are only given to the members of that particular project area.

(3) Personal Area: Members have their own personal area where one can store all kind of work related information and restrict access privileges.

The two core features of the web-applications are as:

(1) Advanced Search: SharePoint 2010 Search service [7] is connected with all of three web applications. Users are able to find information quickly and easily using advanced filters.

(2) Security: Permission levels are enhanced at multiple levels especially at Project Area, Personal Area and file level.

Figure 1 depicts the current structure associated with KMSI.

1) Data Libraries Area:

One may wish to publish a particular content on the IMMPETUS intranet whether it is individual documents, web pages, theses, publications, annual reports, images, videos, experimental data or records with strict regulatory requirements being fully managed within KMSI's architecture. The documents library includes many levels of sites and sub-sites, so that one may use the navigation bar to quickly browse to the content that one wants. Data libraries are provided with advance filtering and sorting features. Information stored here is protected and only authorised users are able to read it. An on-line survey area is also created and results can be stored and viewed in an efficient graphical manner. A multimedia library is also provided where images, audio and video files are being stored; media assets such as videos can be watched directly from the library. The optimum supported media file size is 150MB but up to 2 GB of video file can be stored here [8].

Currently, about 2800 different documents and files have been stored on various data libraries; the average size of a document/file is about 2 MB. Figure 2 shows the KMSI 'Data Libraries Area' home-page for browsing the public libraries.

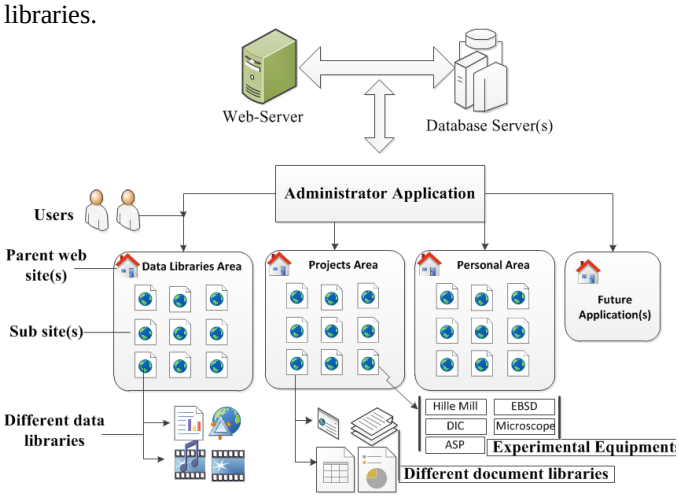


Figure 1. KMSI's configuration model

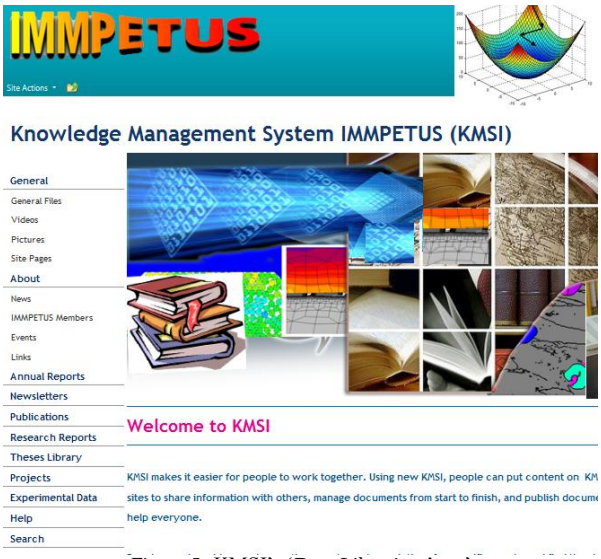


Figure 2. KMSI's 'Data Libraries Area'

2) Projects Area

KMSI has a dedicated website for each IMMPETUS project which acts as a vehicle for effective collaboration and information storage infrastructure for members. Each project area has a facility to store documents, files, experimental data, and organise stored data into different kinds of libraries. Microsoft SharePoint 2010 Team-site [9] templates are used to develop different project workspaces because they are highly flexible and designed to encapsulate collaboration features. Microsoft SharePoint 2010 and Microsoft Office are so strongly integrated that they provide a strong platform for collaboration and sharing [9]. Users having the necessary access rights are able to check-out, check-in, lock or share document(s) among whole of the project team where the permissions have been assigned. Online project management tools are also provided within Projects Area to view the progress of the project tasks. Each library is configured to store up-to 25versions of a file; the maximum file size which one can upload is restricted to 250MB. Figure 3 depicts the structure associated with Projects Area.

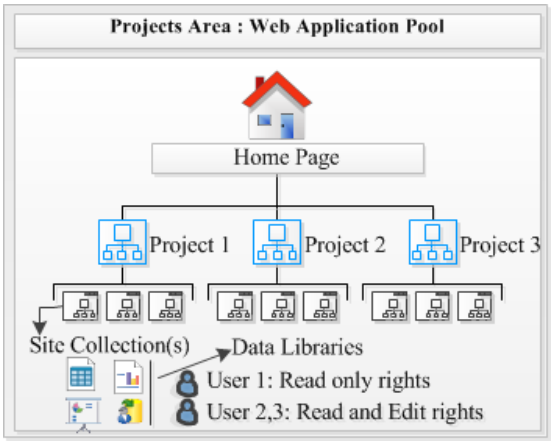


Figure 3. Projects Area structure

Five core data libraries are provided to store and share information in a 'Project Area' as follows:
a) PDF Files: This data library is built to store all PDF files which one can also share with other members.
b) Project Files: All types of project related files can be saved here such as Matlab, Abaqus, and MSC Nastran etc.
c) Presentations / Other files: This library is to store all presentations and power point slides. Users are also able to maintain versions of each file stored here.
d) Multimedia: Users are able to store all project related pictures and videos here.
e) Experimental data: This is a vital part of the Project Area where members are able to save records and data generated from experiments. Each project may have one or more records library e.g. Hille Mill, ASP (Arbitrary Strain Path), TMC (Thermo Mechanical Compression), and DIC (Digital Image Co-relation) etc. Figure 4 shows an interface to store experimental data onto KMSI server.

3) Personal Area

Personal web areas are designed for IMMPETUS academic staff, research associates and PhD students. This site provides a central location to manage and store one's documents, content, images, and useful information. 'Personal Area' serves as a point of contact for other users within IMMPETUS to find some information about the user, user skills, and user interests [10].

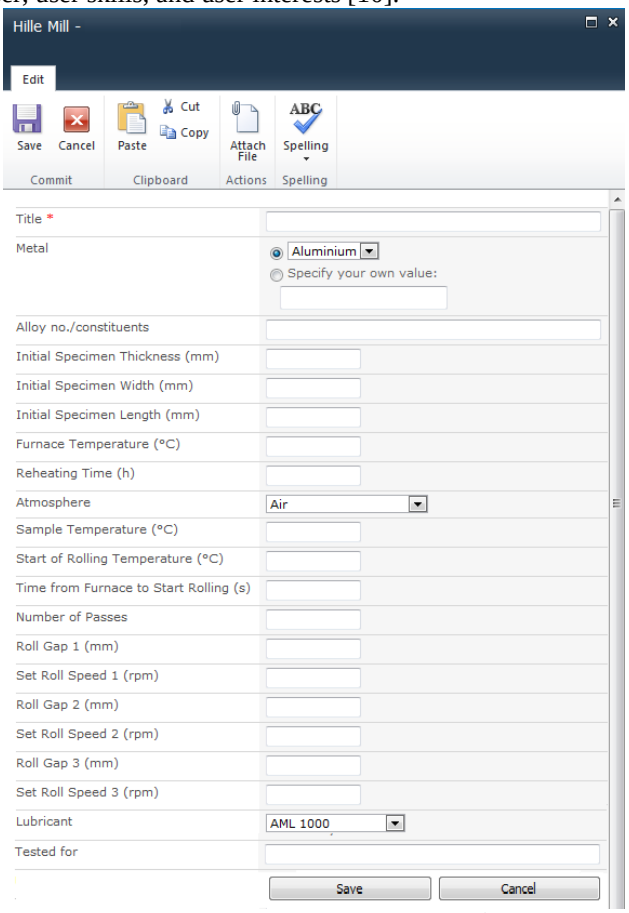


Figure 4. Hille Mill experimental data storage form

One can customise the information and the content on his/her personal web area. The user can manage (save/edit/delete) documents, lists, and images within his/her web area. Personal web area also facilitates users to organize and get their information when and wherever it is needed using a secure connection over the internet.

4) Advanced Search

A previous study [11] found the following: *(1) "Typical employees spend an average of 3.5 hours per week trying to find information but not finding it." (2) They spend 3.0 hours more recreating information they know exists, but that they simply cannot locate*".

KMSI provides an intuitive and flexible user interface, improved relevance, and the ability to search unstructured and structured information such as databases, all public information across the KMSI portal especially focusing on project sites, data libraries and experimental records. Advanced search filters help narrowing the search results to find the desired content quickly. Information can be refined using web area, author, date, document type (pdf, doc, docx, xls etc.), or using tags associated with the content. Figure 4 shows the search results returned by the system against the keyword 'steel'.

5) Security

A set of permissions can be granted [12] to KMSI members on a securable object such as a project site, document library, experimental data rig, folder, item, or document. Permission levels enable the user to assign a set of permissions to other KMSI members so that they can perform specific actions on their project area.
The following permission levels are granted:

- **Full Control:** This permission level contains all permissions (Administrator rights).
- **Area Members:** Can add, edit, and delete items in existing lists and document libraries. Assigned to those who are members of that project area.
- **Read:** Read-only access to the KMSI, users with this permission level can view items and pages, open items, and documents. This permission level is by default assigned to all IMMPETUS members within 'Data Libraries Area'.
- **File Level Permission:** Each site contains additional securable objects that have a particular position in the site hierarchy, and user can also set different levels of permission on various folders, files and contents. Figure 6 illustrates file level permissions for different users.

 ISBN: 978-1-61208-162-5

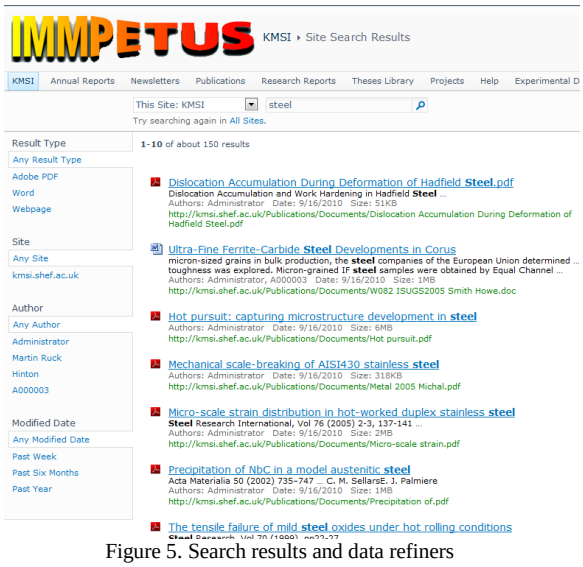


Figure 5. Search results and data refiners

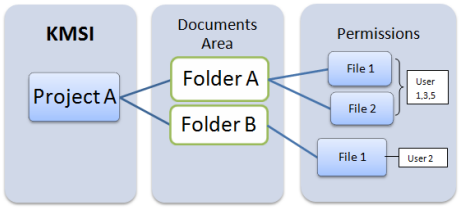


Figure 6. File level permissions for four different users.

IV. CONCLUSION

IT tools and technologies, such as Microsoft SharePoint 2010, are providing robust platform which can be used to develop the Knowledge Management Systems as per one's requirements. Our system (KMSI), which is based on Microsoft SharePoint 2010 technology, helps users gather information in a holistic fashion, regardless of the type of file or information. It also helps users to find, manage and keep track of the updated information in an efficient manner.

KMSI provides IMMPETUS members the necessary tools to manage their documents, research papers, and experimental records in an efficient way. Even more importantly, all this information is easily shared between users, such as project teams, Departments, and/or IMMPETUS members. The experimental data storage facility is improved and data can be easily exported to comma separated values (CSV) or Microsoft Excel format.

We were required to disseminate the required information to subsribed users whenever it is added onto the public libraries or in some project areas. Information is normally filtered and processed using 'managed meta data' [13], content types or using file attributes which help the system to identify information and re-route the information to the appropriate data libray or to some 'Project Area'.

We are having some difficulties to incorporate experimental data directly from some experimental equipment. It is also required to crawl inside some content like Matlab and Abaqus data-files; we are still doubtful how iFilters [14] help to dig inside those files. In future, we are also willing to add virtual experimental equipment inside the KMSI to perform experiments remotely using various data models which is a challenging and brainstorming job.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Engineering and Physical Sciences Research Council for their financial support under grant Ref. No. EP/E063497/1.

REFERENCES

[1] IMMPETUS, "About", www.immpetus.group.shef.ac.uk, (Last Accessed 18 July 2011)

[2] Maryam Alavi and Dorothy E. Leidner, "Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues," MIS Quarterly, Vol. 25, No. 1 (Mar., 2001), pp. 107-136.

[3] Bill English, Briand Elderman and Mark Ferraz, Microsoft SharePoint 2010 Administrator's Companion, 1st ed., Microsoft Press, (2011), pp. 439-440.

[4] Toby Bell, Karen M. Shegda, Mark R. Gilbert and Kenneth Chin, "Magic Quadrant for Enterprise Content Management", 16 November 2010, http://www.gartner.com/technology/media-products/reprints/microsoft/vol14/article8/article8.html, (Last Accessed 12 August 2011)

[5] Boris Pluskowski, Dynamic Knowledge Systems, http://www.imaginatik.com/site/pdfs/WP-0602-1%20Dynamic%20Knowledge%20Systems.pdf, "unpublished", (Last Accessed 12 August 2011)

[6] Bill English, Briand Elderman and Mark Ferraz, Microsoft SharePoint 2010 Administrator's Companion, 1st ed., Microsoft Press, (2011), pp. 397-398.

[7] Microsoft, Enterprise Search, http://technet.microsoft.com/en-us/enterprisesearch/default.aspx , (Last Accessed 12 August 2011)

[8] Microsoft Technet, "SharePoint Server 2010 capacity management", http://technet.microsoft.com/en-us/library/cc262787.aspx, (Last Accessed 12 August 2011)

[9] Microsoft Technet, "Collaboration site planning", May 12, 2010, http://technet.microsoft.com/en-gb/library/cc262388.aspx, (Last Accessed 12 August 2011)

[10] Microsoft Technet, "My Site Overview" , http://technet.microsoft.com/en-us/library/ff382643.aspx, (Last Accessed 18 July 2011)

[11] Bill English, Briand Elderman and Mark Ferraz, Microsoft SharePoint 2010 Administrator's Companion, 1st ed., Microsoft Press, (2011), pp. 442-443

[12] Microsoft Technet, "Plan site permissions", December 16, 2010, http://technet.microsoft.com/en-us/library/cc262778.aspx, (Last Accessed 12 August 2011)

[13] Microsoft Technet, "Managed metadata", http://technet.microsoft.com/en-gb/library/ee424402.aspx , (Last Accessed 12 August 2011)

[14] Microsoft Technet, "File types and IFilters reference", November 11, 2010, http://technet.microsoft.com/en-us/library/gg405170.aspx, (Last Accessed 12 August 2011)

 ISBN: 978-1-61208-162-5

Optimising Parameters for ASKNet: A Large Scale Semantic Knowledge Network Creation System

Brian Harrington, Simon Kempner
University of Oxford Department of Computer Science
Keble College Oxford
Oxford, United Kingdom
Email: brian.harrington@cs.ox.ac.uk
Email: simon.kempner@keble.oxon.org

***Abstract*—ASKNet is a system for automatically constructing semantic knowledge networks from natural language text. ASKNet uses existing natural language processing tools to extract entities and relations from text, and then through a combination of lexical information and a novel use of spreading activation, combines that information into a large scale semantic knowledge network. The ASKNet system is large, and quite complex. Historically, users of the system have had to rely on a combination of intuition and empirical evaluation of small sample networks in order to obtain reasonable settings for the various system parameters. In this paper, we develop a testing harness and gold standard that allow us to use simple machine learning methods to find optimal settings for all of the system's parameters. This system also aids future development of internal system algorithms, and can be adapted easily to novel domains.**

Keywords-Semantic Networks; Natural Language Processing; Spreading Activation; Knowledge Networks; ASKNet; Parameter Optimisation

I. INTRODUCTION

ASKNet is a system for automatically generating large scale semantic resources using information derived from natural language texts. Using a combination of existing natural language processing tools and a novel application of spreading activation, ASKNet builds semantic networks representing the information contained within a text, and then maps that information onto a larger network representing the sum of its world knowledge.

ASKNet has been in development since 2005, and has been shown to produce good results on a variety of tasks, such as Semantic Relatedness [1], [2], and conceptual knowledge acquisition [3]. The integrated semantic nature of ASKNet also makes it ideal for information management and knowledge discovery [4], [5].

Large systems such as ASKNet necessarily have various parameters which must be optimised in order to obtain the best possible results from the system as a whole. During the development of ASKNet, the parameters controlling elements such as the spreading activation and lexical matching were set by the developers based on their own intuition and unit testing. While the system was being refined, small test networks were built to help developers find appropriate values for these parameters, but due to the large scale nature of the project, it was not feasible for any developer to manually configure all of the parameters. Thus, the values were always set to very rough approximations, and even when running tests on new data sets, the system's configuration was often left in the same state as it had been for previous experiments [6].

This paper details the development of a "gold standard" data set, and testing harness for ASKNet, and the use of an evolutionary based hill-climbing algorithm. The combination of these tools allows us to automatically and find optimal settings for parameters. We then use these improved parameters to repeat a previously published experiment, and find an improvement in both precision and running time.

II. ASKNET

ASKNet uses a combination of natural language processing tools such as the C&C parser [7], and the semantic analysis tool Boxer [8] in order to produce discourse representation structures. These structures are then converted into semantic network fragments as seen in Figure 1. The network fragments are based on an entity relationship paradigm, with nesting to allow entities and relations to be combined to form concepts, which can in turn be combined to form structures of increasing complexity (See Figure 2).

Once the semantic network fragment has been created for a piece of text, ASKNet then uses a Spreading Activation based algorithm [4] in order to determine the appropriate mappings between nodes in the fragment and nodes in the global knowledge network.

A. Spreading Activation

In order to integrate the semantic network fragments into the larger knowledge network, ASKNet uses the *update algorithm*, which is based on the psycholinguistic principles of Spreading Activation [9]. Spreading activation works by considering ASKNet networks as having similar properties to neural networks. By placing an amount of activation in a node, and allowing that node to fire, it can spread the

 ISBN: 978-1-61208-162-5

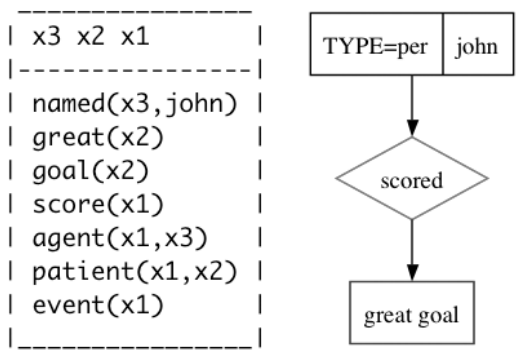


Figure 1. Boxer output, and corresponding semantic network fragment for the sentence "John scored a great goal".

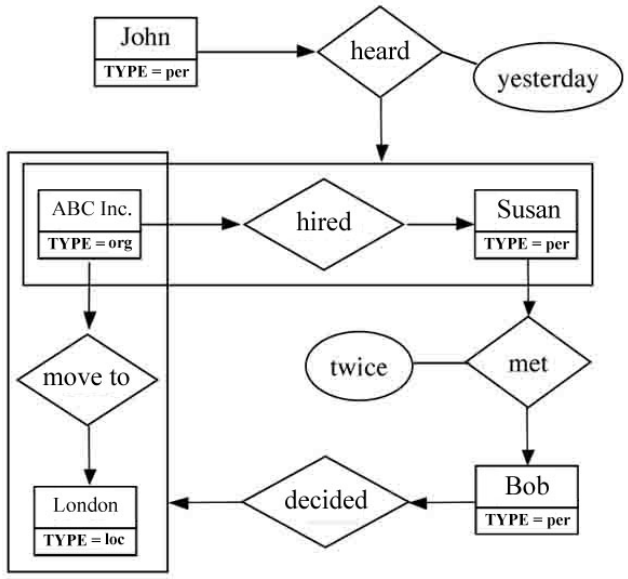


Figure 2. A sample ASKNet network, showing simple elements combining to form more complex concepts.

activation to its neighbouring nodes, the amount passing to each neighbour being relative to the strength of the relation connecting them.

In the update algorithm, ASKNet first uses lexical similarity to get base mapping scores for all named entity node pairs. A small amount of activation is then placed in a *source* node in the network fragment. The activation is allowed to spread through the fragment, settling on various fragment nodes dependant on their relatedness to the source node. The current mappings are then used to transfer that activation from the fragment nodes to corresponding nodes in the main knowledge network, the amount being transferred being dependent on the current mapping score, and the main network is then allowed to fire. The amount of activation received at the end of this process by the main network *target* nodes will determine the update to the (*source*,*target*) mapping score.

Figure 3 shows an example of the update algorithm in progress. An initial mapping score will be created between the pairs (`bu`,`georgebush`) and (`bu`,`johnbush`) based on their lexical similarity (string similarity + named entity type). In order to improve these scores, `bu` is selected as the source node, and given activation which will spread to `go` and `wh` dependant on the strength of the "beat" and "to" relationships. The activation from these nodes will be sent to `whitehouse`, `algore` and `gorevidal` respectively based on their relative mapping score. The main network will then be able to fire, resulting in activation filtering to the `georgebush` node, while the `johnbush` node receives no activation. Thus, the mapping score for (`bu`,`georgebush`) will increase, and the score for (`bu`,`johnbush`) will decrease. This process will continue until the scores reach a stable state, or cross a threshold at which time the nodes will be mapped together.

The update algorithm allows ASKNet to integrate information from a variety of sources into a single cohesive semantic network. Spreading activation has the advantage of being localised, and thus relatively efficient, while at the same time taking into account the relative strength of multiple paths of varying lengths that may connect node pairs.

B. Previous Evaluation

In a previous paper [6], ASKNet networks were created from documents provided in the 2006 Document Understanding Conference [10]. Each of the 5 networks, each containing information from 25 documents was then given to 3 judges in order to evaluate their quality. The judges were asked to evaluate the paths between each pair of named entities in the network, and mark the path as either "entirely correct" (all entities, mappings and relations were correct), or "incorrect" (there was an error of any type).

In the original experiment, it was found that manual evaluation of the entire network was impractical, and therefore the evaluation was performed on the "core" of each network. The core being defined by the named entities that were mentioned in more than 10% of the documents, and the paths connecting them. One of the network cores is shown in Figure 4.

The human evaluators found that an average of 79.1% of the paths were correct, with a Kappa Coefficient [11] of 0.69 indicating a high level of agreement between evaluators. A breakdown of the scores is provided in Table I.

In order to evaluate the work presented in this paper, we will recreate a portion of this experiment.

 ISBN: 978-1-61208-162-5

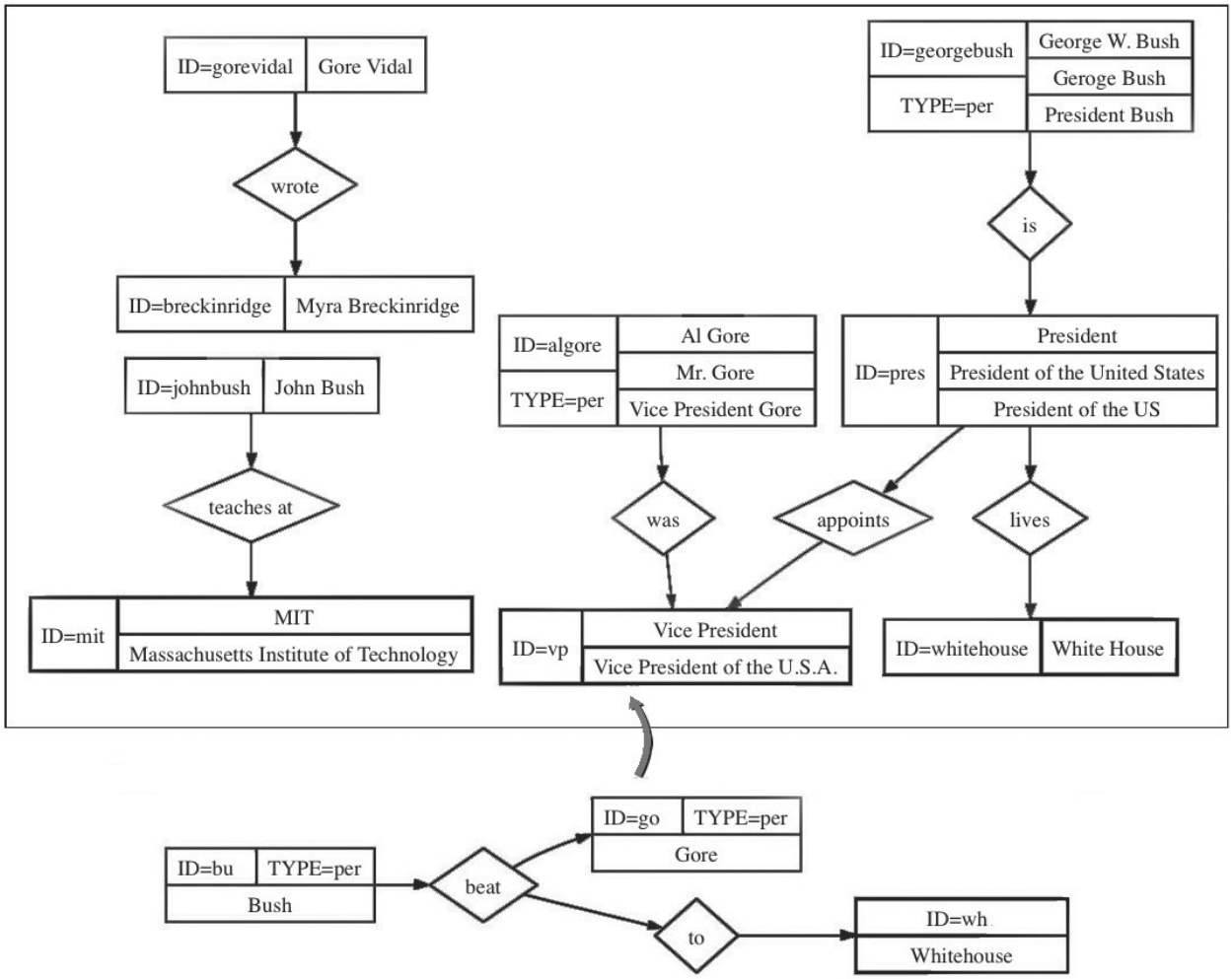


Figure 3. An example ASKNet semantic network fragment being added to a knowledge network relating to U.S. politics

Topic	Eval 1	Eval 2	Eval 3	Avg
Elian Gonzalez	88.2%	70.1%	75.0%	77.6%
Galileo Probe	82.6%	87.0%	91.3%	87.0%
Viruses	68.4%	73.7%	73.7%	71.9%
Vladimir Putin	90.3%	82.8%	94.7%	89.9%
West Bank	68.2%	77.3%	70.0%	72.3%
Average Precision:				79.1%

Table I
EVALUATION RESULTS FOR THE 2008 EXPERIMENT

III. ESTIMATING PARAMETERS

A. Developing a Gold Standard

The first step in developing a method for automated parameter refinement is to create a gold standard evaluation. In order to build such a resource, 742 lines of text were processed from the BBC News Business edition RSS feed (*http://feeds.bbci.co.uk/news/business/rss.xml?edition=int*). A GUI tool (See Figure 5) was created that allowed users to select, for each potential mapping that ASKNet would consider, whether that mapping was correct. 2 evaluators were asked to complete the mappings using the tools, and in the case of disagreements, a third evaluator was asked to break ties. A total of 1306 mappings were produced in under 30 minutes per evaluator, with an inter-rater Kappa Coefficient of 0.989, indicating an extremely high level of agreement.

B. An Evolutionary Hill Climbing Search

For our experiments, we attempted to optimise the settings for 5 parameters.

All parameters were initially set to their default values provided by the system developers. Then ASKNet was run on the BBC data, and the mappings were compared against those in the gold standard. A weighted harmonic mean was used to calculate an F-Score of 0.436. A weighting of $\frac{3}{4}precision$ to $\frac{1}{4}recall$ was chosen to emphasise the

 ISBN: 978-1-61208-162-5

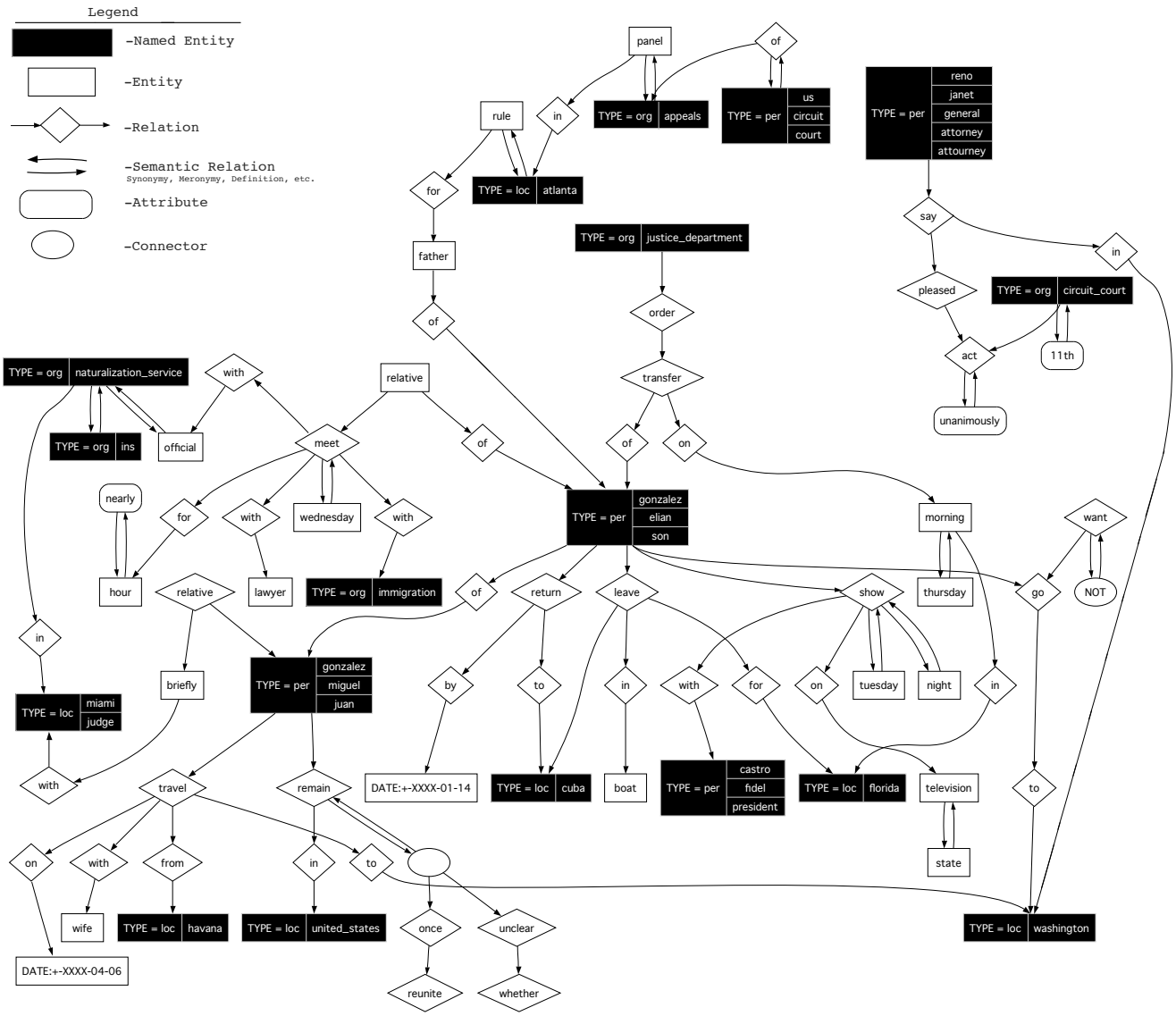


Figure 4. The core of the ASKNet network containing information on the Elian Gonzalez custody trial.

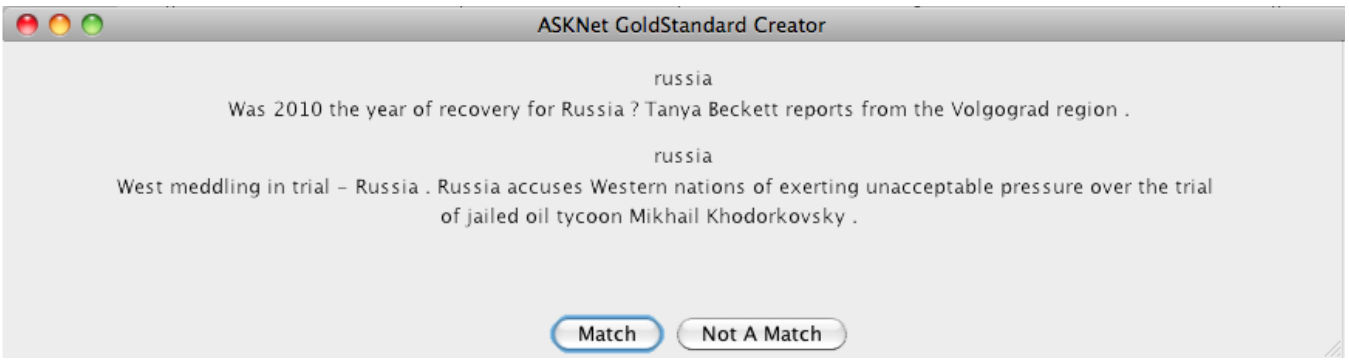


Figure 5. The user interface of the gold standard maker.

mapThreshold	The mapping score level above which we map a node pair together
initialActivation	The amount of activation initially given to the source node
iterations	The number of iterations of the firing algorithm to be run
signalAttenuation	The amount of signal that is lost with each firing (controls the maximum distance that activation can spread from its source)
firingThreshold	The minimum activation required to cause a node to fire

importance of precision in this context.

An evolutionary hill climbing algorithm based on machine learning techniques was then implemented in which parameters were adjusted in turn by small delta values, repeating the gold standard test until the F-Score was optimised for a local maximum. Once a local maximum was reached, a parameter was selected at random to be "mutated" to a random value.

This hill climbing search was allowed to run for 2 hours on a 2.4GHz processor, and eventually resulted in a maximum weighted F-Score of 0.510. The parameters that achieved these values were then saved for use in the experiment.

IV. Evaluation

In order to evaluate the performance improvement that our optimised parameters generated, ASKNet was run on the same data set as was used in the 2008 experiment detailed in Section II-B. However, since we are only concerned with improving the mapping, and have not affected the parsing or semantic analysis, we chose to modify the experiment to focus on the precision of the mappings, as opposed to the overall network.

The experiment was repeated, but with evaluators only being asked to judge whether the mappings were correct. They were asked to evaluate each named entity in the network, and provide a score of "correct" (The entity corresponds to a single real world entity, and all instances of that entity have been correctly mapped onto a single node) or "incorrect" (Two or more real world entities have been mapped to a single node, or one real world entity has been split between multiple network nodes).

The experiment was first run with the original settings, yielding an overall precision of 71.6%. It should be made clear that these results are lower than those presented in Table I, as the new experiment is focusing solely on the mappings, which is the most difficult element of network creation, and thus would have produced a higher proportion of errors than the parsing and semantic analysis phase.

The improved settings were then tried, yielding a result of 79.5%. An improvement of nearly 8%. This means that simply optimising 5 of the parameters across the system removed nearly 8% of the errors made by the system. While this may not seem like a vast improvement at first, in a large scale network such as those built in previous experiments

Topic	Eval1	Eval2	Eval3	Avg
Elian Gonzalez	61.3%	58.0%	64.6%	61.3%
Galileo Probe	78.2%	72.3%	80.1%	76.9%
Viruses	73.5%	68.2%	74.7%	72.1%
Vladimir Putin	81.2%	84.4%	89.0%	84.9%
West Bank	61.2%	62.3%	64.2%	62.6%
Average Precision				71.6%

Table II
EVALUATION RESULTS WITH DEFAULT SETTINGS

Topic	Eval1	Eval2	Eval3	Avg
Elian Gonzalez	70.3%	69.1%	75.2%	71.5%
Galileo Probe	86.4%	78.9%	82.0%	82.4%
Viruses	73.1%	69.3%	72.2%	71.6%
Vladimir Putin	84.4%	88.9%	94.7%	89.3%
West Bank	80.2%	82.1%	85.3%	82.5%
Average Precision				79.5%

Table III
EVALUATION RESULTS WITH OPTIMISED SETTINGS

[3], this could remove tens of thousands of possible errors. In this experiment, the inter-rater Kappa coefficient was 0.72, once again indicating a high level of agreement between all three evaluators, and confirming that the improved score shown is due to an actual improvement in the mappings, and not due to evaluator bias.

V. Conclusion

In this paper, we have developed an automatic system to optimise the parameters of ASKNet using a gold standard annotated by human evaluators, and a hill climbing algorithm with genetic mutations. With the parameters optimised by this system we are able to improve the precision of the system's mapping ability, one of the core functionalities of ASKNet, by almost 8%, as shown by a manual evaluation.

These techniques can be useful both in improving the quality of the networks produced by ASKNet, but also allow researchers to efficiently tune parameters to new data sources and types of information. In order to build an ASKNet network on a new type of information, such as scientific text or narratives, it is only necessary for researchers to develop a new gold standard markup, using the tools already provided, and use the same hill-climbing algorithm to find optimised parameters.

In this experiment, we only chose to optimise the 5 most important parameters in ASKNet. However, these techniques could be used to evaluate more fundamental changes, such as re-designing of the underlying algorithms and data structures. By providing ASKNet developers with a simple, efficient, automated tool to adjust settings, and a gold standard against which to test, developers can efficiently evaluate changes they are making to ASKNet without having to rely on intuition, or undergo the relatively time-consuming task of developing large scale networks.

 ISBN: 978-1-61208-162-5

REFERENCES

[1] B. Harrington, "A semantic network approach to measuring relatedness," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, 2010.

[2] P.-R. Wojtinnek and S. Pulman, "Semantic relatedness from automatically generated semantic networks," in *Proceedings of the 9th International Conference on Computational Semantics*, Oxford, UK, 2011.

[3] P.-R. Wojtinnek, B. Harrington, S. Rudolph, and S. Pulman, "Conceptual knowledge acquisition using automatically generated largescale semantic networks," in *Proceedings of the 18th International Conference on Conceptual Structures*, Kuching, Sarawak, Malaysia, 2010.

[4] B. Harrington and S. Clark, "Asknet: Automated semantic knowledge network," in *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI'07)*, Vancouver, Canada, 2007, pp. 889–894.

[5] B. Harrington, "Managing uncertainty, importance and differing world-views in asknet semantic networks," in *Proceedings of the fourth IEEE International Conference on Semantic Computing*, Pittsburgh PA, USA, 2010.

[6] B. Harrington and S. Clark, "Asknet: Creating and evaluating large scale integrated semantic networks," *International Journal of Semantic Computing*, vol. 2, no. 3, pp. 343–364, 2009.

[7] S. Clark and J. R. Curran, "Wide-coverage efficient statistical parsing with CCG and log-linear models," *Computational Linguistics*, vol. 33, no. 4, pp. 493–552, 2007.

[8] J. Bos, S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier, "Wide-coverage semantic representations from a CCG parser," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, Geneva, Switzerland, 2004, pp. 1240–1246.

[9] G. Salton and C. Buckley, "On the use of spreading activation methods in automatic information retrieval," in *Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval*. New York, NY, USA: ACM Press, 1988, pp. 147 – 160.

[10] H. T. Dang, "Overview of duc 2006," in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, New York NY, USA, 2006.

[11] J. Carletta, "Assessing agreement on classification tasks: the Kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

 ISBN: 978-1-61208-162-5

Semi-Automated Semantic Annotation for Semantic Advertising Networks

Aseel A. S. Addawood
Department of Information Science
College of Computer & Information Sciences,
Imam Mohammad bin Saud University
Riyadh, Saudi Arabia
Addawood@ccis.imamu.edu.sa

Lilac A. E. Al-Safadi
Department of Information Technology
College of Computer & Information Sciences,
King Saud University
Riyadh, Saudi Arabia
lalsafadi@ksu.edu.sa

***Abstract*—In this work, we present a semi-automatic semantic annotation system which is designed to link publisher entries to an existing Ontology and instances. It assists in categorizing the content of Web sites and associating advertisements with publishers. Semantic annotation can enhance information retrieval and improve interoperability. Since manual annotation is inefficient, Automatic or semi-automatic annotation makes the process of annotation fast and easy. The suggested system is integrated in the publisher's registration platform of the semantic advertising network and serves to facilitate registration, semantic annotation and information utilization in Web sites. This system is part of a semantic advertising network prototype called Lexeme.**

Keywords-semantic web; semi-automatic annotation; Advertising Networks; RDF.

I. INTRODUCTION

Advertising networks or ad networks are companies that connect potential advertising Web sites with potential advertisers [5]. Ad networks have made advertising on the Web very easy. Advertisers pay Web site owners ("Publishers") to allow ads to be shown on their sites. The Publisher is an individual or corporation responsible for the distribution of digital publications who will be using the ad network to make a profit from facilitating easy and dynamic publishing of ads on his Web site.

The reliance of ad networks on the keywords (in the content) without an accurate interpretation of the context of the page results in a display of irrelevant and unappealing ads on a Web page [1]. The Semantic Web is a technology that can be utilized by publishers to analyze the meanings behind a word or words. It will help to place ads in prime web locations for the sole purpose of reaching their targeted consumers [2]. We envision an algorithm that contains not only information instructing machines as to what ads to display, but also structured data which make machines understand what ads have been displayed. Such structured information that can be read and understood by computers [3] is the key. It enables machine-to-machine exchange and automated processing in a way that computers can understand [4]. In a time of mass content creation, improving ad placement through more optimized, findable content ushers in a new era of Semantic technology . It delivers the right message to the right user. The Semantic advertising networks combine the desirable features of both advertising networks and the Semantic Web. Semantic documents, Web sites and ads are generally written in the Resource Description Framework (RDF) and Web Ontology Language (OWL) languages [4]. Currently, only a few semantic documents exist on the Internet.

The challenge addressed by this paper is related to automating the provision of semantic structure to publishing Web sites. The semantic structure is provided to individual pieces of information in the Web site and interlinks these pieces with semantic relations. This results in a meaningful organization of content.

The paper is organized as follows. In the next section, we present the semantic representation of Web site content, and in Section 3, we show the related work. In Section 4, we discuss the proposed semantic annotation system. In Section 5, we discuss what technology has been used in developing the prototype. In Section 6, we show the experiment results. The conclusion and future work are given in Section 7.

II. SEMANTIC REPRESENTATION OF WEB SITE CONTENT

In this section, we focus on annotating Web site content with semantic representations. Semantic annotation helps in effectively matching Web site content with relevant resources.

Currently, a document might be one page filled by text. The data itself are not structured in a way that can be interpreted by a computer. There is no complex logic or reasoning concerning the data; there are only simple keyword-matching algorithms. At this stage, it is necessary to establish a relation among the data so that it can be considered a semantic web.

However, the next generation of the Web, the Semantic Web, makes machines more intelligent as a result of better algorithms used to process data on the Web. Web 3.0 is about data that is connected and capable of being reassembled on demand. This reassembly of data and the reorganization of data pieces is a central factor of Web 3.0. [6].

 ISBN: 978-1-61208-162-5

The resulting intelligence in the structure and format of the data yields a richer relationship and linking infrastructure of data on the Web. The Semantic Web specifications, in particular RDF and OWL, are the only technology specifications that were purpose-built for use as a metadata language entirely dedicated to describing and linking data of all sorts at Web scale [7]. Therefore, the Semantic Web introduces a logical language that human programmers can use to inform computers of the relationships among data. It pursues an important goal of creating a new form of Web content that is meaningful to computers.

Ontology is an explicit formal specification of the terms in the domain and relations among them [9]. Ontology can be used to define the underlying semantics of the Web site content through the semantic annotation system, as illustrated in Figure 1 below. Semantic annotation helps in linking Ontology with Web site content for an efficient and easy embedding of semantics. The annotation results are stored in an RDF metadata store.

III. RELATED WORK

This paper is motivated by the need for adding metadata to existing web pages in an efficient way and establishing relationships among the data. There have already been certain researches conducted which are connected with our mechanisms for semi-automatic semantic annotation, described in Section 4. Our attempt is to demonstrate the application of semantic annotation of publishing Web sites to serve advertising network applications. The suggested annotation system supports an integrated environment with the registration of publishing Web sites in a semantic advertising network system. In addition, it supports document-annotation consistency and separate annotation storage. It automatically links salient terms in a Web site to relevant ontological instances/classes and their properties. It uses simple lexicon-based parsing and linguistic rules to identify instances.

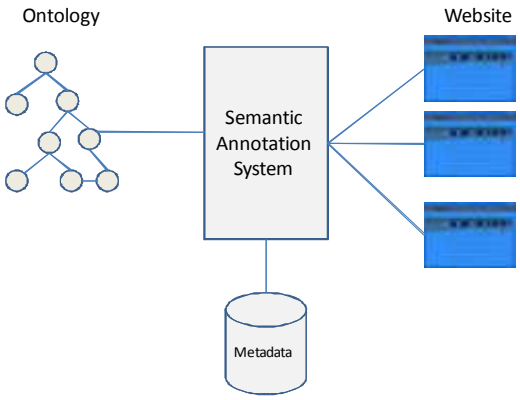


Figure1. Linking Web sites and Ontology by Semantic Annotation

Erdmann et al. [12] describe their approach to finite state technologies and support of lexical acquisition, and semantic tagging through them. The work coincides with our approach and uses the concepts that are stored on the Ontology. The source Websites have been manually annotated to explicitly represent the semantics of their contents. It introduces a proprietary extension to HTML that is compatible with common web browsers. Since there is a huge amount of relevant information for most communities, manual annotation is burdensome, and it is an impractical solution.

Blythe et al.'s [13] ACE system enables users to enter a free text into a parser. Then it compares the free text with the Ontology for term replacement. However, the ACE system cannot annotate the whole article. It only accepts user annotations as short statements of free text. The system is designed to be robust, allowing partial formalizations of the annotation and not relying on a successful parse of the user's input.

Steffen et al. [14] have been developing a tool, OntoAnnotate, that allows usage of domain-specific Ontologies for easy annotation of HTML documents and creation of meta data by (semi-automatic) annotating web pages. Starting from their Ontology-based annotation environment in OntoAnnotate, they have collected experiences in an actual evaluation study.

Liu et al. [15] propose a semi-automatic annotation system, which assists users in annotating textual web data and manages the terms defined by the user. The system uses OWL to describe semantic web data and annotate them. It happens in two ways: using a manual annotator with the help of the user, or an automatic annotator using the KMP algorithm. The work uses string matching to identify classes and neglect properties and relationships.

Most of the works on Ontology-based semantic annotation have been developed based on manual semantic annotation. With the huge amount of information on the Web and the upswing of the Semantic Web, there comes an urgent need for automating the process of adding semantic annotation to existing web pages.

IV. SEMANTIC ANNOTATION SYSTEM

This section illustrates how the semantic annotation system works. It gives an example, and then describes the suggested architecture of the system and the different components that compose it.

A. Example Scenario

The proposed system links Web site registration to an existing education advertising domain Ontology (EAO) [9] by semi-automatic semantic annotation. A Web site is described as a set of concepts followed by a set of properties. When the publisher registers his Web site, he has to enter the URL of the Web page that will host the

 ISBN: 978-1-61208-162-5

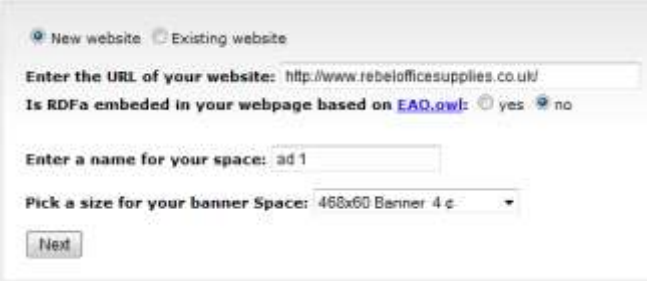


Figure 2. The registration page of the publisher

advertisements . For example, in our case, the Web site [17] is an educational Web site, and it was one of our samples. In addition, the publisher has the choice of entering a URL that has an RDFa embedded in his Web page (i.e., Semantic Web site), as shown in Figure 2.

Annotation starts with parsing the content of the registered Web site for extracting salient terms. This is automated by natural language processing of the Web site content.

Our system links extracted concepts to Ontology entries and suggests a number of terms that may describe the Web page, depending on the content of the publisher Web site and the Ontology. The publisher has to choose a number of terms that describe his Web page the best. In our example, the system will suggest the concepts "Color," "Ink," "Computer" " Paper," and "Printer" as shown in Figure 3. As long as we are suggesting a semi-automated annotation system, the system expects verification of the results suggested by the parser.

Then, the system gives a list of properties expressed in a natural language, which is suggested by the Ontology content. In our case it corresponds to the "Paper" and "Computer," namely: "is manufactured by," is dimensioned," "is colored," "is priced for," as shown in Figure 4. The publisher selects a property and then assigns to it a specific value. In our example, the price of the paper was set to "10$" and the computer manufactured by "Dell" and priced for "1500$". That represents the value of the selected property "is priced for" and the property" is manufactured by."

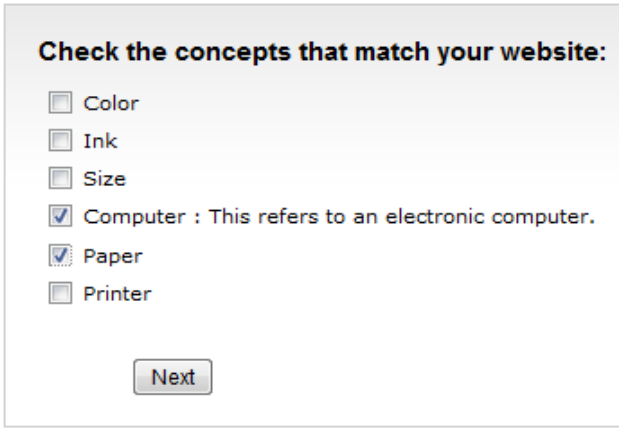


Figure 3. The concepts that match publisher Web site the best

Figure 4. Properties of the selected concept

The system is self-learning. In cases in which some related instances or concepts are not defined in the Ontology, only the administrator may add a suitable instance or concept. The administrator may enter a new relation for an existing concept as shown in Figure 5 below.

Furthermore, if the concepts found on the publisher Web page cannot be matched with the Ontology concepts, the system suggests terms that have a similar meaning. For example, if it finds "Pen" or "Writing Instrument" in the Web page, then it will refer to the term "Pen." Figure 6 shows an XML file storing all the concepts that are in the Ontology and their synonyms. The XML file can be edited through administration access only. It is used to give each term a score that measures the relevancy to the concept that is stored in the Ontology. The relevancy score is a measuring function that is conducted by the system to help in matching the terms found in the Web pages to our Ontology concepts. The administrator can add and/or edit the concepts in the XML file and calculate the score of the relevancy.

 ISBN: 978-1-61208-162-5

Figure 5. Adding a new concept to the Ontology

```
<keyword name="Pen">
<term name="pen" score="100"/>
<term name="writing instrument" score="100"/>
</keyword>
```

Figure 6. Synonym of the term "Pen"

```
<keyword name="Notebook">
<term name="lined Notebook" score="90"/>
<term name="unlined Notebook" score="90"/>
<term name="notebook" score="100"/>
<term name="stationery" score="100"/>
</keyword>
```

Figure 7. Synonyms of the term "Notebook"

Also, if the system finds a type of concept, the system will suggest the concept to the publisher. For example, as shown in Figure 7, the system will suggest the term "Notebook" if it finds one of these terms: "lined Notebook", "unlined Notebook", "notebook" or "stationery".

B. Proposed Architecture

Figure 8 illustrates the architecture of our suggested system, which helps the publisher to add semantic description to his Web page in a semi-automated way. The publisher first adds his Web site metadata along with the URL through the Web site registration page. The Term Extractor is an automatic tool used to extract terms that best describe concepts in the Web page. We have used the Porter stemming algorithm (or 'Porter stemmer'). It is a process for removing the more common morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems [8].

The Concept Mapper links extracted concepts to Ontology entries. The component automatically suggests related instances and saves verified annotations in the metadata store. Further, the system suggests relationships defined between instances by finding related instances from the Ontology. Finally, the publisher is provided with a list of properties associated with each instance and asks for supplying values. The set of concepts, relationships

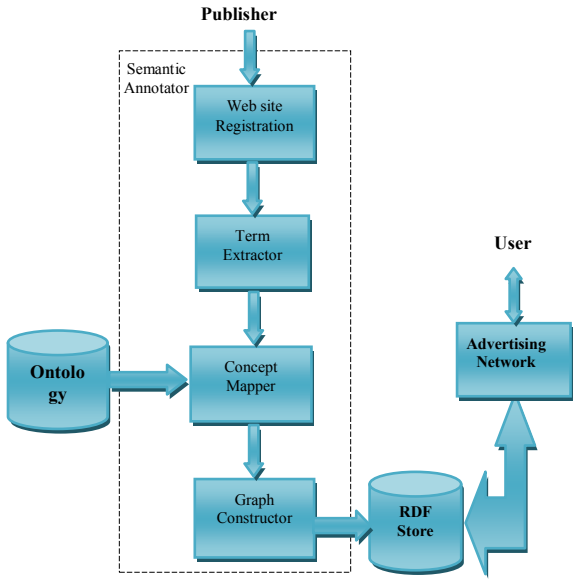


Figure 8. The proposed architecture for the semi-automated semantic annotation system

```
<rdf:RDF
 xmlns:eao=http://www.lexeme-ads.com/EAO.owl/
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
<rdf:Description rdf:about="http://www.lexeme-
ads.com/#Computer">
<eao:hasManufacturer
xml:lang="en">Dell</eao:hasManufacturer>
<eao:hasPrice xml:lang="en">1500</eao:hasPrice>
<rdf:type rdf:resource="http://www.lexeme-
ads.com/EAO.owl/Computer"/> </rdf:Description>
<rdf:Description rdf:about="http://www.lexeme-
ads.com/#Paper">
<eao:hasPrice xml:lang="en">10$</eao:hasPrice>
<rdf:type rdf:resource="http://www.lexeme-
ads.com/EAO.owl/Paper"/>
</rdf:Description>
</rdf:RDF>
```

Figure 9. A fragment of the RDF Representation of the www.rebelofficesupplies.co.uk content

and properties describing the Web site semantic content is referred to as the RDF model.

The RDF model will be converted into an RDF graph through the Graph constructor. An RDF graph is set of RDF triples (subject, verb, and object). Triples are the basis of information representation. Figure 9 above shows part of the RDF code that has been generated for the Web site [17].

V. IMPLEMENTATION

To develop the semi-automated system, we used Jena's APIs [18] as our semantic framework. We also developed

 ISBN: 978-1-61208-162-5

an inference engine that links both advertisers to publishers and vice versa. In addition, we asked a few experts to gather

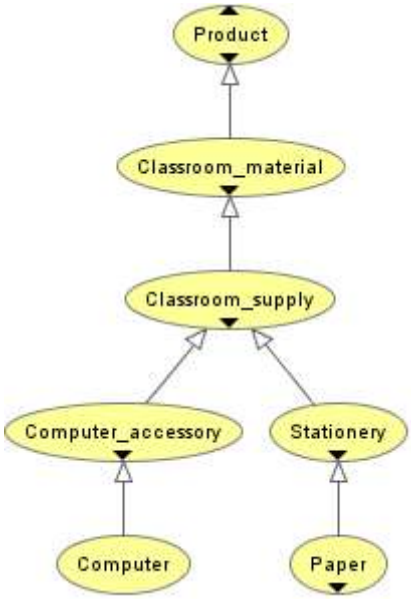


Figure 10. The relationship among the entities

some samples that they considered to be related to the educational domain. We then made sure that they were 100% strictly XHTML pages.

Failure to find a proper Ontology for putting relevant ads on WebPages stimulated us to create our own product. By developing a custom-made Ontology, we were able to define countless education specialties, such as publications, school supplies, and writing instruments, that had not been present in any educational Ontology at the time. To build the Ontology, we used the Protégé editor [19]. Figure 10 shows a snapshot of the structure of the Ontology of the above scenario.

VI. EXPERIMENT RESULTS

The system provides indexed RDF matches and metadata matches. In this section, we test the system using two evaluation methods. First, we assess the RDF files that have been generated by the system for each website using the W3C validation service [20] to validate the RDF files. Experiments have been carried out using ten educational Web sites, the advertisement collection is approximately 50 ads, and all generated RDF files have been validated successfully through this tool.

Second, we evaluate the performance of the system using Precision and Recall. We do not intend to suggest a sophisticated semantic annotation system. Instead, we provide a simple demonstration of semantic annotation in advertising networks.

Semantic match in the system retrieves publishing web sites relevant to a submitted ad of interest. A simple algorithm is outlined below.

1. Get and conceptualize the ad.
2. Find relevant publisher's web sites by matching the ad RDF against the RDF repository of publisher's web sites.
3. Retrieve the metadata of the relevant Web sites.
4. Place the ad in the ad place defined for ad displays in the publisher's Web site.
5. Capture the number of clicks on the ad and add to web site metadata.
6. Direct the visitor to the corresponding ad home page.

Jena's matching method is used to match the semantic content of both the ad and the publishing web sites.

The following experiment aims at testing our generated RDFs for the publishers' Web sites against the stored RDFs of the ads. The matching process for both approaches is the same.

The system will match the RDF that has just been created with the stored RDFs in the database and try to find a match. The matching process will depend on the number of triples in the RDF graph that matches with the Web site graph. If it finds a matching triple or a partial matching, it will put the matching advertisement in the Web site space. Figure 11 shows part of the code that is used for matching. The system uses the getSubject() and getPredicate() methods that are impeded in Jena to match the two graphs depending on their subject and predicate.

```
public static int Match(Graph g1, Graph g2)
{
   Triple T1,T2;
   int count;
   ExtendedIterator iter1, iter2;
   count = 0;

    /* if the two graphs are isomorphic; stop the loop and the ad will be
placed in this website. the count will be = -1 */

    if(g1.isIsomorphicWith(g2))
     return -1;
     iter1 = g1.find(Triple.ANY); //returns an Iterator for all the triples in
the graph
  while(iter1.hasNext())
    {
     iter2 = g2.find(Triple.ANY);
      T1 = (Triple) iter1.next(); // T is a triple from the ad graph
      while(iter2.hasNext())
      {
         T2 = (Triple) iter2.next();
         if (T1.getSubject().equals(T2.getSubject()) &&
T1.getPredicate().equals(T2.getPredicate()))
          {
            count++;
          } } }
return count;
}
```

Figure 11. Part of the matching code

```
<rdf:RDF
 xmlns:eao=http://www.lexeme-ads.com/EAO.owl/
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >

  <rdf:Description rdf:about=http://lexeme-ads.com/#Computer>
<eao:hasDimension xml:lang="en"> 11 inches</eao:hasDimension>
<rdf:type rdf:resource="http://lexeme-ads.com/EAO.owl/Computer"/>
  </rdf:Description>
</rdf:RDF>
```

Figure 12. A fragment of the RDF Representation of the ad

Figure 13. Displaying the result of the match

The RDF file that is shown for the matching ad is shown in Figure 12.

Among the set of advertisements returned by the system to be displayed on the publisher Web site, we selected the result shown in Figure 13, from an advertisement for the Dell company, which was placed in the matched website.

We have used the Precision and the Recall to compute the matching Web sites with relevant ad rates. The following table shows the Precision and Recall rates for each website. Also, it shows the F-score, which is defined as the harmonic mean of Precision and Recall to reflect the actual performance of the system.

Precision is a measure of correctness. As Table 1 shows, the Precision rates are mostly 100% because the system only retrieves the ads that have matching RDF graphs with the website graphs. If the graph does not match, the system will discard the advertisement as a choice and test another advertisement graph.

On the other hand, Recall is a measure of completeness. It shows the rate of retrieving all the relevant ads, as shown in the table below, where it ranges from 30% to

TABLE 1. PRECISION, RECALL AND F-MEASURE FOR A SEMANTIC ADVERTISING NETWORK

website No.	Precision(%)	Recall(%)	F-measure(%)
1	75	30.8	44
2	100	100	100
3	100	100	100
4	100	75	85.7
5	100	75	85.7
6	100	55	70.9
7	100	88	93.6
8	100	55	70.9
9	100	33.3	49.6
10	100	37.5	54.5

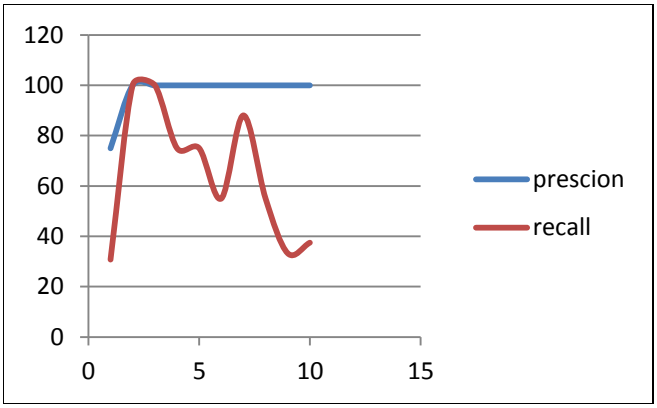


Figure 14. the relation between the Precision and the Recall

100%. The reason for that is that when the system chooses one ad and tries to find another ad that matches, it will encounter a relevant ad; however, sometimes that relevant ad will have fewer matching triples then the original ad. The system will not consider it as retrieved.

As a result of a greater Precision rate, the Recall rate is decreased, as shown in Figure 14, and as the Precision rate becomes higher, the number of relevant ads that are retrieved is lower.

If we combine the two metrics, Precision and Recall, we get their harmonic mean, known as the F-measure. This is a measure of a test's accuracy. As shown in Table 1, the F-measure ranges from 44% to 100%. The F-measure average is 75.5%. This average is considered high and it points to the high accuracy of the system. Considering the experiment results, we believe semantic advertising networks have considerable potential.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented a prototype of a semi-automated semantic annotation system for semantic advertising networks. This helps the publisher in describing

his Web page using semantic technology and Ontology. We have demonstrated and proven how current advertising methods can be improved. The introduction of Semantic Technology is essential for reaping larger financial gains. It is important for analyzing the meanings between the lines in order to place ads in prime web-locations for the sole purpose of reaching their targeted consumers. With this system, there will no longer be the Web 2.0 emphasis on factors such as keyword matching. The emergence of Web 3.0 means tapping into more important elements such as understanding the context of Web pages, and latent or hidden connections amid these contexts. It serves to discover relationships among concepts and ideas.

In the future, a significant amount of work should be done. We will try to provide the publisher with the privilege of adding/editing concepts to the Ontology by an administration approval that matches his preference. We will support the RDF standard for representing metadata on the web, representing both Ontologies and generated annotated facts in RDFs. This standard will make annotated facts reusable and machine processable on the web [10].

ACKNOWLEDGMENT

Our thanks to Arwa Al-Tameem, Ghada Abuguyan, May Abu Melah, Nora Al-Zaid, Nouf Al-Najran, and Nadeen Al-Abdullatif for developing the prototype.

We are thankful to Dr. Muhammad Abulaish working at the Center of Excellence in Information Assurance, King Saud University for his review and guidance.

REFERENCES

[1] W. D. Wells, S. Moriarty, and J. Burnett, "Advertising: principles and practice", 7th ed. New York: Prentice Hall, 2005.

[2] M. Strckland, "Guide to Semantic Web: Creating more relvent ads",[Online]. Available: https//www.threeminds.organic.com/ [Accessed: Sept. 9, 2011].

[3] T. Wilson, "How semantic web works", [Online]. Available: http//computer.howstuffworks.com. [Accessed Sept. 9, 2011].

[4] D. Allemang, and J. Hendler, "Semantic web for the working ontologist: effective modeling in RDFS and OWL". Burlington, Massachusetts: Morgan Kaufmann, 2008, pp. 1-40.

[5] L. Al-Safadi, A. Al-Dawood, and N. Abdulateef, "Lexeme: An Ontology-based semantic advertising networks". Journal of Computing, 2(9), 2010, pp.1-5.

[6] M. Marshall, "The semantic web & its implications on search marketing",[Online].Available:http//www.searchenginejournal .com/. [Accessed Sept. 9, 2011].

[7] J. Davies and R. Struder, "Semantic web technology: trends and research in ontology system-based systems". Chichester, Wiley, 2006, pp.29-50.

[8] "The Porter Stemming Algorithm",[Online]. Available: http://tartarus.org/~martin/PorterStemmer [Accessed Sept. 9, 2011].

[9] L. Al-Safadi, N. Abdulateef, "Educational advertising ontology: a domain-dependent ontology for semantic advertising networks". Journal of Computer Sciences 6(9), 2010, pp.1-8.

[10] R. Benjamins, D. Fensel, and S. Decker. 1999. "KA2: building ontologies for the internet: a midterm report". International Journal of Human Computer Studies, 51(3):687.

[11] Lassila, O. and Swick, R. (1999). Resource description framework (RDF) model and syntax specification. Technical report, W3C. W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax.

[12] M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab, "From manual to semi-automatic semantic annotation: about ontology-based text annotation tools". In P. Buitelaar & K. Hasida (eds). Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, pp.3-7.

[13] Blythe J. and Gil Y. "Incremental formalization of document annotations through ontology-based paraphrasing". In Proceedings of the 13th International World Wide Web Conference (New York, New York, May 2004), pp. 455-461.

[14] Steffen Staab, Alexander Maedche and Siegfried Handschuh. "An annotation framework for the semantic web". In P. Buitelaar & K. Hasida (eds). in proceedings of the first workshop on multimedia annotation, pp. 1-4.

[15] C.-H. Liu, H.-C. Chen, J.-L. Jain, and J.-Y. Chen. "Semi-automatic annotation system for owl-based semantic search". International Conference on Complex, Intelligent and Software Intensive Systems (2009 IEEE), pp.1-5.

[16] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: requirements and a survey of the state of the art," Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol.4(1), 2006, pp. 14–28.

[17] "Rebel Office Supplies", [Online].Available: www.rebelofficesupplies.co.uk. [Accessed Sept. 18, 2011].

[18] "Jena – A Semantic Web Framework for Java", [Online].Available: jena.sourceforge.net. [Accessed Sept. 18, 2011].

[19] "Protégé", [Online]. Available: protege.stanford.edu. [Accessed Sept. 18, 2011].

[20] "W3C Validation Service", [Online]. Available: www.w3.org/RDF/Validator/. [Accessed Sept. 18, 2011].

 ISBN: 978-1-61208-162-5

Mining Cross-document Relationships from Text

Petr Knoth
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
p.knoth@open.ac.uk

Zdenek Zdrahal
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
z.zdrahal@open.ac.uk

***Abstract*—The paper argues that automatic link generation and typing methods are needed to find and maintain cross-document links in large and growing textual collections. Such links are important to organise information and to support search and navigation. We present an experimental study on mining cross-document links from a collection of 5000 documents. We identify a set of link types and show that the value of semantic similarity can be used as a distinguishing indicator.**

Keywords-text mining; automatic link generation and typing; semantic similarity; digital libraries.

I. INTRODUCTION

There has been a significant research effort in the area of modelling cross-document relationships. These include various semantic relations at the discourse level ranging from mere similarity of topics presented in two documents to the assertion that one document elaborates/contradicts the ideas described in another one. Enriching document collections by inter-document relationships provides the means for better organising fragmented information. In practice, this would improve the browsing, the navigation and the discovery of important information resources. However, the current cross-document relationship modelling approaches rely on human annotators and therefore they do not scale-up. So far, little work has seriously addressed the limitations of manual identification of cross-document relationships in large and constantly growing repositories. In this paper, we argue that automatic link discovery and typing methods can be used to bridge this gap. In this paper, the concept of *link* refers to a semantic connection between two segments of text, such as two documents or paragraphs, at the discourse level and should not be confused with the Semantic Web representation known as Open Linked Data, which is an approach for publishing data and their relations using RDF triples.

This work is based on the following hypotheses:

(a) Cross-document links can be generated automatically using semantic similarity as one of the criteria.
(b) The value of semantic similarity is related to the link type.

The paper brings the following contributions:

- We provide evidence and argue why automatic link generation is necessary for the creation and maintenance of typed relationships, especially in scholarly databases and encyclopedias, and why it cannot be easily substituted by social tagging or crowdsourcing approaches.
- We elaborate the abovementioned hypotheses, especially (b), for a selection of link types.
- We present a simple experiment for mining link types motivated by the results previously reported in [1].

The rest of the paper is organized as follows. In Section II, the role automatic link generation methods can play in automatically analyzing large text collections is introduced. Related work in the areas of semantic web tools for discourse modeling, automatic link generation and link typing is discussed in Section III. In Section IV, we argue why automatic link generation is needed and cannot be substituted by crowdsourcing approaches. An experimental link typing study is presented in Section V. Finally, the paper is concluded in Section VI.

II. AUTOMATIC MINING OF CROSS-DOCUMENT LINKS

The automatic link generation task can be defined as follows: Let S and T be collections of documents, denoting sources and targets respectively. Let $s \in S$ and $t \in T$ be lexical units of possibly different granularity. For example, s and t can be the whole documents, paragraphs, sentences or even noun phrases. The goal is to find a binary relation $\rho \subseteq S \times T$ defined in terms of pairs $< s_i, t_j >$ such that all pairs are interpreted by a human evaluator as carrying the same semantic relationship. For example, ρ can be interpreted as *is similar*, *is_the_same*, *expands*, *contradicts*, etc. The relation must satisfy the usual properties, e.g., *is_the_same* is symmetric, transitive and reflexive, *is_similar* is not transitive, *expands* is antisymmetric etc.

Automatic link generation methods have many potential applications. For example, the methods can be used for the interlinking of resources not originally created as hypertext documents, for the maintenance or the discovery of new links in collections growing in size, or to improve navigation in collections with long texts, such as books or newspaper articles. All this makes the automatic mining of cross-document links a very useful technology which could be applied across a number of disciplines including information retrieval, semantic web, user navigation, text summarization and others.

 ISBN: 978-1-61208-162-5

III. Related work

A. Semantic web technology for cross-document relationship modeling

One of the most important areas where cross-document relations play a key role are digital libraries. Nowadays, the activities of researchers and students rely more and more on access to large online repositories using technologies and tools, such as Google Scholar, CiteSeer or PubMed. These systems currently do not provide support for organising, modeling and sharing cross-document relationships. Consequently, their navigation capabilities are limited.

To fill the gap, scientific community invested significant effort into relationship and argument modeling tools. For example, the Mendeley tool [2] allows to discover related research literature, highlight and organise it, annotate relationships to other articles and share them with others. Similar work has been done previously by Uren et al., the ClaiMaker tool described in [3] allows to model and share research debates/discourses across scientific literature. Other work has also focused on relationship and argument visualization [4]. A number of tools have also been developed for specific domains, such as the life sciences [5], [6].

Though the abovementioned studies recognise the potential offered by collaborative tagging, crowdsourcing and sharing, the resulting approaches rely in the end always on human annotators. We claim that there are at least two reasons why such an approach cannot scale-up: (1) The rate of information growth is faster than the resources of the crowds. This issue is further discussed in Section IV. (2) Researchers are usually reluctant to share this type of knowledge, because the skill of analyzing and interpreting papers is the researcher's know-how. This has also been recognised in the tool presented in [3], where sharing is restricted to a selected research community.

B. Link generation

In the 1990s, the main application area for link generation methods were hypertext construction systems [7]. Nowadays, link generation methods for finding related documents have become the de-facto standard. They have been applied in large digital repositories, such as PubMed or the ACM Digital Library, or in search engines including Google Scholar. Generating links pointing to units of a lower granularity than a document has been investigated more recently. The task of such systems is to locate relevant information inside the document instead of only providing a link to the whole document. The Initiative for the Evaluation of XML retrieval (INEX) played an important role in the link generation research by providing evaluation tracks (Link-the-Wiki track) for link generation systems at the granularity of documents as well as at a more fine-grained granularity [8].

Current approaches can be divided into three groups: (1) *link-based* approaches discover new links by exploiting an existing link graph [9], [10], [11]. (2) *semi-structured* approaches try to discover new links using semi-structured information, such as the anchor texts or document titles [12], [13], [14]. (3) *purely content-based* approaches use as an input plain text only. They typically discover related resources by calculating semantic similarity based on document vectors [15], [16], [17], [18]. Some of the mentioned approaches, such as [11], combine multiple methods.

C. Link taxonomies/ontologies and link typing

A pioneering study in link typing has been presented already in 1980s by Randall Trigg [19] who developed a taxonomy of link types. Trigg divided links into two groups - normal (inter-document) links and commentary (cross-document) links. His rich taxonomy of link types enables the specification of judgements on hypertext nodes. With link types, such as *unimportant*, *solved*, *insufficient* or *incoherent* the taxonomy is *content focused* rather than *relation focused* [20]. Another approach is represented by the ScholOnto taxonomy [21] which has been developed with a reference to cognitive coherence relations [22].

An influential study on automatic generation and typing of links has been published in [23]. Allan recognizes that certain cross-document link types (*automatic links*) can be automatically extracted more easily than others. He focuses then on the development of methods for the identification of the automatic link types, involving relations such as *tangent*, *equivalence* or *contrast*.

An unsupervised approach for the recognition of discourse relations has been presented in [24]. The authors show that from a set of adjacent sentences a subset of discourse relations, namely *contrast*, *explanation-evidence*, *condition* and *elaboration* can be recognized with high accuracy. This task is significantly more difficult, but also more interesting, in the cross-document settings. Similar problem has been recently addressed in Radev et. al. [25] who introduced a taxonomy of 18 cross-document rhetorical relationships denoted as Cross-document Structure Theory (CST). In addition, they present the development of an annotated dataset of CST relationships and experiment with the recognition of their subset using machine learning with a varying level of success for different relationships.

IV. Manual annotation and crowdsourcing

Cross-document discourse modeling, i.e., connecting a claim found in one document with a claim found in another one by a semantic relation, such as *contradicts*, is technically identical to the problem of providing metadata that allow to organize resources and information in a logical way. Various social annotation tools for metadata generation available on the Web have become very popular, such as image tagging or rating systems. Most applications that use them are based on the idea that a large number of users can provide in most cases good quality metadata. However, there is a number

 ISBN: 978-1-61208-162-5

of problems where the knowledge of the crowds is not sufficient due to lack of human expertise or theoretical time constraints.

It has been shown [26] that metadata can be divided into three distinct groups with respect to the nature of information they are describing. (1) Metadata describing the content of a resource (2) Metadata classifying a resource using a taxonomy (3) Metadata connecting two resources usually by a semantic relation. While provision of type (1) metadata can be done by humans for large text collections in a reasonable time, the provision of type (2) metadata is problematic and type (3) metadata cannot be manually acquired even in moderately large collections. The reason is that the number of possible connections explodes quadratically with respect to the number of resources and as a result people are unable to keep track of all the relevant available information. The problem appears to be particularly significant in quickly growing collections with many contributing authors.

A tempting approach to resolve this problem is by increasing the number of people who contribute to the collection maintenance, for example, by creating discourse links and then sharing the results with others. Shum and Fergusson expect that this will result in a user-generated web of meaningfully connected annotations which can be visualized, filtered and searched for patterns in ways that are impossible at present [27]. In reality, this approach can be successful only in very limited domains, it certainly does not scale-up unless automatic link generation and typing tools assist in the annotation and the maintenance process. In addition to that, human annotators have been previously found inconsistent in carrying out this task [28].

To provide an example, let us consider Wikipedia, which is today perhaps the largest collection of documents containing user created links and at the same time maintained by a very large community of users (about 250,000 contributing users). Even though Wikipedia contains currently 3,433,587 articles, it is still very small in comparison to all information available on the Web. While in Wikipedia content is typically linked from an anchor (concept) to the whole article (description of the concept), the situation is more complex in other domains, such as in scholarly databases. In Wikipedia, there can be only one page describing a concept whereas in scholarly databases there can be a large and growing number of papers discussing the same topic. The growth of Wikipedia in terms of new articles has already started to decrease and it is predicted that this trend is going to continue in the future. An opposite trend can be expected with scholarly literature.

Even though the problem of linking information less complex in Wikipedia than in scholarly databases and even though the community is very large, the maintenance of Wikipedia is problematic and automatic tools are desperately needed. For example, it has been noted in [29] that the effort necessary for the maintenance of the information on Wikipedia is not directly proportional to the amount of information stored, but rises faster than linearly with the amount of information being added.

V. USING SEMANTIC SIMILARITY FOR LINK TYPING

We have previously studied the relation between links authored by people and links predicted by automatic link generation methods [1], namely using semantic similarity measures on document vectors directly extracted from text. The results indicate that semantic similarity is strongly correlated to the way people link content. In this paper, we are extending this work by investigating the qualitative properties of links. As a test-bed we are using articles selected from Wikipedia. For our experiments, this dataset has the following advantages:

- A large number of good quality articles forming a network of cross-references created and agreed by a sufficiently large community of Wikipedia contributors.
- Articles connected by a single *unspecified* link type. However, the link may represent different semantic relationships.
- A suitable initial test-bed. Only a limited set of discourse relations are present in Wikipedia at the article level. As a consequence, we do not investigate relations, such as disagreement or contradiction that typically do not appear at this level.

The correlation has been measured on a collection of 5,000 Wikipedia articles in categories containing the phrase "United Kingdom". This required the calculation of semantic similarity (in this case *cosine similarity* calculated on *tfidf* document vectors) for $\frac{5{,}000^2}{2} - 5{,}000 = 12{,}495{,}000$ pairs of documents and the extraction of all 120,602 links between these articles created by Wikipedia authors.

A. Linked-pair likelihood

A central concept of our study is the quantity called *linked-pair likelihood* introduced in [1] which is the probability that a pair of documents is connected by a manually created link, calculated as $lpr = \frac{|\text{links}|}{|\text{document pairs}|}$. Figure 1 shows lpr calculated for groups of document pairs at different intervals of semantic similarity. It can be observed that linked-pair likelihood strongly correlates with the value of sematic similarity (this provides an answer to hypothesis (a) in the introduction), however the direction of the correlation is in the right part of the graph quite unexpected. The correlation has been tested for statistical significance with a positive result for p-value well beyond $p < 0.001$ for both Spearman's rank and Pearson correlation coefficients. This indicates that high similarity value is not necessarily a good predictor for the existence of a link. The detail of this experiement can be found in [1].

 ISBN: 978-1-61208-162-5

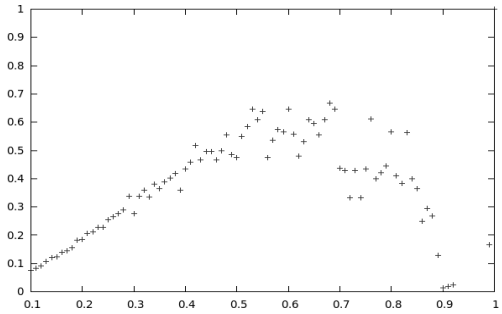


Figure 1. The linked-pair likelihood (y-axis) with respect to the cosine similarity (x-axis) [1].

B. Using semantic similarity for relation typing

The results presented in the previous section provoke a number of questions. Perhaps the two most interesting are:

(1) Why is the curve in Figure 1 not monotonically increasing which would mean the more semantic similar the more likely to be linked?
(2) As content can be linked for various reasons, are there any qualitative differences between linked documents with different value of semantic similarity?

A possible explanation for question (1) is that people create links between related documents that provide new information and therefore do not link nearly identical content. Regarding question (2), we hypothesize that the value of semantic similarity might be used in link type identification, i.e., the reasons for linking articles with different values of semantic similarity are also different. Investigation of these two questions provides answers to hypothesis (b) presented in the introduction.

C. Relations of interest and their representation

In our experiment, we have decided to use four discourse link types building on the classification provided by [23] as we hypothesize that the value of semantic similarity might be a useful distinctive factor. The sampled document pairs were classified to the following types: *tangent*, *similarity/equivalence*, *expansion*, *aggregate*. Examples of these link types are depicted in Table I. The description of these link types is as follows:

Tangent links represent according to [23] links which relate topics in an unusual manner, for example, a link from a document about "Clouds" to one about Georgia O'Keeffe (who painted a mural entitled *Clouds*). In our work tangent links are associated to document pairs that are related in a useful, but relatively marginal way, typically there is a single piece of information that justifies the relationship of the documents.

Expansion link type is attached to a link which starts at a discussion of a topic and has as its destination a more detailed discussion of the same topic.

Title 1	Title 2	Link type	Description
Jack McConnell	Scottish Qualifications Authority	tangent	The first article mentions that the Scottish Labour politician Jack McConnell appointed a new board for the Scottish Qualifications Authority (SQA) and introduced significant changes to the way the agency worked.
Social Democratic Party (UK)	David Owen	expansion	David Owen was was one of the founders of the British Social Democratic Party (SDP) and led the SDP from 1983 to 1987 and the re-formed SDP from 1988 to 1990. The first article mentions David Owen a number of times.
Senior Railcard	Family and Friends Railcard	similarity/ equivalence	Both articles describe the history of railcards introduced by British Rail. Articles clearly describe two semantically related concepts.
Statutory Instruments of the UK, 1996	Statutory Instruments of the UK, 1996 (3001-4000)	aggregate	The first article contains the other as its part.

Table I
EXAMPLE LINK TYPES

Similarity/equivalence links represent related and strongly-related discussions of the same topic.

Aggregate links are those which group together several related documents. According to Allan [23], aggregate links may in fact have several destinations, allowing the destination documents to be treated as a whole when desirable. In our work, only pairs of documents are considered and thus aggregate links are assigned to document pairs when the first article contains significant parts of the second article.

The only discourse link types from Allan's taxonomy that we did not use for classification are *comparison* and *contrast* links. Contrast and comparison is in a Wiki typically handled either explicitly in the text, e.g., *"The invasion of Iraq was particularly controversial, as it attracted widespread public opposition and 139 of Blair's MPs opposed it."* or it is part of the elaboration, revision and refinement process of the article. This obviously reduces the number of discourse relationships we can identify to those mentioned above. We also assume that two contrasting text segments would often be represented by similar term-document vectors and therefore the value of semantic similarity would not provide sufficient information.

D. Results

To answer the questions defined in Section V-B, we have carried out a study that investigates the characteristics of link pairs at different similarity levels. The interval $[0.1, 1]$ of semantic similarity depicted in Figure 1 has been divided

 ISBN: 978-1-61208-162-5

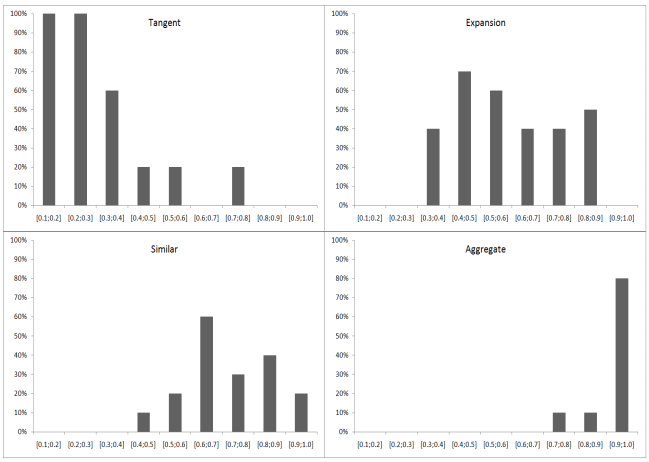


Figure 2. The frequency of different link types with respect to semantic similarity of document pairs

into 9 intervals of even width. As a case study, 10 article pairs from each interval[1] between which a link was created by Wikipedia users were randomly sampled and they were assessed by a human investigator and classified. An evaluation environment was created to allow the investigator to see the articles next to each other and to easily compare them. The investigator was asked to inspect both articles, to assign exactly one of the four relationships of interest and to provide a brief justification for the decision. The document pairs were presented to the investigator in a random order and the investigator was during the evaluation not aware of the calculated value of semantic similarity associated with the article pairs. The evaluation and classification of one pair took from 5 to 20 minutes. The whole manual evaluation took about 19 hours.

Overall, 37% of article pairs were classified as *tangent*, 36% as *expansion*, 20% as *similar* and 7% as *aggregate*. The results of the evaluation are presented in Figure 2. The figure shows the frequency of different link types in all the 9 selected intervals.

We have found that in the lower levels of semantic similarity $[0.1, 0.3]$ most of the links were classified under the tangent link type. At higher levels of similarity the proportion of the tangent link types decreases. Only very few links were classified as tangent when the similarity of the document pair was high.

Expansion links start to appear at similarity higher than 0.3. At the similarity level of $0.3 - 0.4$ the proportion of the expansion links is roughly the same as the proportion of tangent links. The highest proportion of expansion links is present in the semantic similarity interval of $0.4 - 0.6$ where the value of similarity seems to be quite a distinctive factor from the similarity link types. At higher similarity values, the proportion of expansion links drops and similar link types appear.

Most of the similar/equivalence links types are present in the interval $0.6, 0.9$. The proportion of this link type is in this region approximately 40%. It seems that it is hard to distinguish them in this interval from the expansion links solely based on the similarity value. When semantic similarity reaches the value of 0.9, it is possible to see aggregate link types that are characteristic by a large value of similarity.

Overall, this confirms that the value of semantic similarity is a useful factor characterizing up to certain extent the type of the semantic relationship which provides answer to the second question reported in Section V-B. We have also observed from this experiment and Figure 1 that people link most often document pairs of the expansion and tangent types, even though the tangent type is in absolute numbers the most frequent link type. People link less likely document pairs providing similar, equivalent or even duplicate content.

The value of semantic similarity is just one criterion which is useful for the detection of certain link types, but has not been used in link typing previously. We expect that robust link typing systems should be developed by combining a number of strategies. We are aware that the value of semantic similarity as presented in this example is unable to make distinctions about certain link types, such as the *prerequisite* link type, nor it can be used to determine the direction of the link. Other text characteristics perhaps combined with external knowledge should be used for this purpose.

VI. Conclusion

We have shown that automatic link generation and typing systems are needed in order to provide scalable solutions to document interlinking in large text collections. We argued that cross-document relations cannot be simply produced by the "Social Web" using crowdsourcing methods. However, the automatically identified relations can be confirmed or rejected using social tagging and both approaches can work in synergy.

We have presented an experimental study that shows that the value of semantic similarity is a useful indicator that can help to identify link types. We assume that more similar indicators exist and their combination would improve the accuracy of link typing. In our study, we have used Wikipedia as a source of textual document. This choice alowed us to simplify the problem by considering only a limited set of cross-document relations. In the future, we plan to perform similar experiments on data from scholarly databases that provide more complex and challenging environment for link generation and typing. In addition, we plan to work with lexical units of a lower granularity, such as paragraphs, sentences and noun phrases. This will help us to better understand the characteristics of cross-document relationships with the aim to find distinctive features for

[1]Only 5 article pairs were sampled from the interval [0.9,1.0] due to lack of data in this region.

 ISBN: 978-1-61208-162-5

various relationship types. This should enable the building of automated and scalable tools for automatic link generation and typing capable of supporting various reasoning and navigation tasks outlined in the beginning of this paper.

REFERENCES

[1] Knoth, P., Novotny, J., Zdrahal, Z.: Automatic generation of inter-passage links based on semantic similarity. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, Coling 2010 Organizing Committee (2010) pp. 590–598

[2] Henning, V., Reichelt, J.: Mendeley - A Last.fm For Research? In: 2008 IEEE Fourth International Conference on eScience, IEEE (2008) pp. 327–328

[3] Uren, V., Shum, S.B., Li, G., Domingue, J., Motta, E.: Scholarly publishing and argument in hyperspace. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, ACM (2003) pp. 244–250

[4] Shum, S.B.: Cohere: Towards web 2. 0 argumentation (2008)

[5] Burns, G.A.P.C., Cheng, W.C.: Tools for knowledge acquisition within the neuroscholar system and their application to anatomical tract-tracing data. Journal of Biomedical Discovery and Collaboration **1** (2006) pp. 10+

[6] Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The swan biomedical discourse ontology. J. of Biomedical Informatics **41** (2008) pp. 739–751

[7] Wilkinson, R., Smeaton, A.F.: Automatic link generation. ACM Computing Surveys **31** (1999)

[8] Huang, W.C., Geva, S., Trotman, A.: Overview of the inex 2009 link the wiki track. (2009)

[9] Itakura, K.Y., Clarke, C.L.A.: University of waterloo at inex 2008: Adhoc, book, and link-the-wiki tracks. [30] 132–139

[10] Jenkinson, D., Leung, K.C., Trotman, A.: Wikisearching and wikilinking. [30] 374–388

[11] Lu, W., Liu, D., Fu, Z.: Csir at inex 2008 link-the-wiki track. [30] pp. 389–394

[12] Geva, S.: Gpx: Ad-hoc queries and automated link discovery in the wikipedia. In Fuhr, N., Kamps, J., Lalmas, M., Trotman, A., eds.: INEX. Volume 4862 of Lecture Notes in Computer Science., Springer (2007) pp. 404–416

[13] Dopichaj, P., Skusa, A., Heß, A.: Stealing anchors to link the wiki. [30] pp. 343–353

[14] Granitzer, M., Seifert, C., Zechner, M.: Context based wikipedia linking. [30] pp. 354–365

[15] Allan, J.: Building hypertext using information retrieval. Inf. Process. Manage. **33** (1997) pp. 145–159

[16] Zeng, J., Bloniarz, P.A.: From keywords to links: an automatic approach. Information Technology: Coding and Computing, International Conference on **1** (2004) pp. 283

[17] Zhang, J., Kamps, J.: A content-based link detection approach using the vector space model. [30] pp. 395–400

[18] He, J.: Link detection with wikipedia. [30] pp. 366–373

[19] Trigg, R.: A Network-Based Approach to Text Handling for the Online Scientific Community. PhD thesis (1983)

[20] Mancini, C.: Cinematic Hypertext: Investigating a New Paradigm. (2005)

[21] Shum, S.B., Motta, E., Domingue, J.: Scholonto: an ontology-based digital library server for research documents and discourse. International Journal on Digital Libraries **3** (2000) pp. 237–248

[22] Sanders, T., Spooren, W.P.M., Noordman, L.G.M.: Coherence relations in a cognitive theory of discourse representation. Cognitive Linguistics (1993)

[23] Allan, J.: Automatic hypertext link typing. In: HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext, New York, NY, USA, ACM (1996) pp. 42–52

[24] Marcu, D., Echihabi, A.: An unsupervised approach to recognizing discourse relations. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2002) pp. 368–375

[25] Radev, D.R., Zhang, Z., Otterbacher, J.: Cross-document relationship classification for text summarization. Unpublished paper (2008)

[26] Knoth, P.: Semantic annotation of multilingual learning objects based on a domain ontology (2009)

[27] Shum, S.B., Ferguson, R.: Towards a social learning space for open educational resources. In: OpenED2010: Seventh Annual Open Education Conference. (2010)

[28] Ellis, D., Furner-Hines, J., Willett, P.: On the measurement of inter-linker consistency and retrieval effectiveness in hypertext databases. In: SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, Springer-Verlag New York, Inc. (1994) pp. 51–60

[29] Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in wikipedia. In: CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (2007) pp. 453–462

[30] Geva, S., Kamps, J., Trotman, A., eds.: Advances in Focused Retrieval, 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and Selected Papers. In Geva, S., Kamps, J., Trotman, A., eds.: INEX. Volume 5631 of Lecture Notes in Computer Science., Springer (2009)

 ISBN: 978-1-61208-162-5

Analyzing the Use of Word Graphs for Abstractive Text Summarization

Elena Lloret
Dept. of Software and Computing Systems
University of Alicante
Apdo. de correos, 99
E-03080, Alicante, Spain
Email: elloret@dlsi.ua.es

Manuel Palomar
Dept. of Software and Computing Systems
University of Alicante
Apdo. de correos, 99
E-03080, Alicante, Spain
Email: mpalomar@dlsi.ua.es

***Abstract*—This paper focuses on abstractive text summarization. Our aim is to explore to what extent new sentences generated employing a word graph-based method (which either compress or merge information) are suitable for producing abstracts. Moreover, in order to decide which of the new sentences should be included in the abstractive summary, an extractive text summarization approach is developed (i.e., COMPENDIUM), so that the most relevant abstractive sentences can be selected and extracted. As shown by the results obtained, this task is very challenging. However, preliminary experiments carried out prove that the combination of extractive and abstractive information is a more suitable strategy to adopt towards the generation of abstracts.**

Keywords-Human Language Technologies, automated retrieval and mining, automated content summarization, abstractive techniques, graph-based algorithms.

I. INTRODUCTION

Currently, the necessity of having good systems and tools capable of dealing with all the information available in an efficient and effective manner is crucial to provide users with the specific information they are interested in. In light of this, Text Summarization (TS) is of great help since its main aim is to produce a condensed new text containing a significant portion of the information in the original text(s) [1].

The process of summarization can be divided into three stages [2]: *topic identification*, *topic interpretation* and *summary generation*. Extractive summarization relies on the selection of the most important sentences in order to produce the summary. As a consequence, only carry out the *topic identification* step is carried out. In contrast, abstractive approaches require a more elaborate process, involving sentence compression, information fusion, and/or language generation. In these cases, all the stages of the summarization process are taken into account.

Due to the difficulty associated to the generation of abstracts, most approaches only focus on the first stage (i.e., topic identification), producing extracts as a result [3], [4], [5], [6]. The main problem of extractive summarization, though, concerns the coherence of the resulting summaries, since the sentences contained may not be properly linked, and most of them will suffer from the well-known *dangling anaphora* phenomenon, i.e., when the pronouns in a summary do not refer to their correct antecedent. Consequently, in order to solve these limitations, research into abstractive methods is needed [7], [8], [9].

The aim of this paper is to conduct an analysis of the potentials and limitations of word graphs for generating abstractive summaries. We first propose a method for compressing and merging information based on word graphs, and then we generate summaries from the resulting sentences. This allows us to quantify how feasible it is to produce abstracts directly. The results obtained give clear proof of the difficulty of the task, and the challenges it presents. However, in a preliminary experiment, we show that a more appropriate strategy would be to combine extractive and abstractive information, improving the performance of the resulting summaries considerably.

The remaining of the paper is structured as follows: Section II introduces previous work in abstractive techniques. Section III describes the word graph-based method for compressing and merging sentences. Further on, how abstractive summaries are produced is explained in Section IV. Section V provides all issues concerning the experiments and evaluation. Additionally, Section VI shows a preliminary analysis of two proposed strategies in an attempt to solve the limitations found in the approach. Finally, the conclusions of the paper together with the future work are outlined in Section VII.

II. RELATED WORK

In this section, we explain previous work on recent abstractive summarization, and we stress our novelty with respect to other similar approaches.

An approach for combining different fragments of information that have been extracted from one or more documents is suggested in [10]. From a predefined vocabulary (e.g., *to address*), the algorithm is able to decide which of these expressions is more appropriate for a sentence, depending on the content and the partial abstract generated. Using machine learning techniques and experimenting with different types of classifiers, results showed that the best classifier, based

 ISBN: 978-1-61208-162-5

on summarization features was able to correctly predict 60% of the cases.

Furthermore, sentence compression [11], [12], and sentence fusion [13], [14] are techniques that have also been applied to abstractive summarization. In particular, graph-based algorithms used for such purpose have been proven to be very successful for producing multi-document summaries [15], [16]. On the one hand, regarding sentence fusion, in [15], related sentences are represented by means of dependency graphs, and then the nodes of such graphs are aligned taking into account their structure. Then, Integer Linear Programming [17] is used to generate a new sentence, where irrelevant edges of the graphs are removed, and an optimal sub-tree is found employing structural, syntactic and semantic constraints. On the other hand, for sentence compression, Filippova [16] suggests a method based on word graphs, where the shortest path is computed to obtain a very short summary (only one sentence) from a set of related sentences belonging to different documents.

Liu and Liu [18] attempt to transform an extractive summarization into an abstractive one in the context of meeting summarization by performing sentence compression. Different compression algorithms, such as Integer Programming, Markov Grammars [19] or even human compression were evaluated, with the result that there are certain limitations when using only sentence compression for generating abstracts. With the same idea, Steinberger et al. [20] explore different ways to generate summaries from their representations through their most important sentences. Their aim is to remove unnecessary words from the original sentences, and then use a probabilistic approach to try to reconstruct them. This approach was found to obtain similar results to extractive summarization.

Our research focuses on studying the applicability of a compression and fusion strategy for producing abstractive summaries. We rely on word graphs for representing documents, and we use them to produce single-document abstracts, allowing the algorithm to compress and merge information.

III. Using Word Graphs for Generating New Sentences

In this section, we explain the proposed algorithm based on word graphs for generating new sentences. Such sentences can be either a compressed version of the original one, or a longer sentence containing information from several.

A. Building the Word Graph

A document is represented as a directed weighted graph $DG = (V, E)$, where $V = v_i, v_{i+1}, \ldots, v_{i+n}$ is the set of nodes corresponding to document's words, and $E = e_{i,i+1}, e_{i+1,i+2}, \ldots, e_{i-n,i+n}$ is the set of edges, which consists of adjacency relations between the words. For the implementation we used Python-graph library [21]. Two words are mapped into the same node only if they have the same part of speech by using TreeTagger [22]. It is important to stress that stop words are not mapped together; otherwise, the real meaning of the sentence could be changed when generating the new sentence. In the future, we plan to use semantic knowledge in order to be able to map concepts instead of words.

In addition, we have to define a weighting function $W(e_{i,i+1})$ for each edge, in order to determine how relevant the edge is. The proposed weight takes into account the frequency of occurrence (*FreqRel*) of two words together in the document, as well as the importance of the words themselves, which is determined through their PageRank value (*PR*) [23]. Therefore, the weighting function can be computed according to Formula 1.

$$W(e_{i,i+1}) = \frac{1}{FreqRel(v_i, v_{i+1}) * (PR(v_i) + PR(v_{i+1}))} \quad (1)$$

In Figure 1, a fragment of a graph is shown.

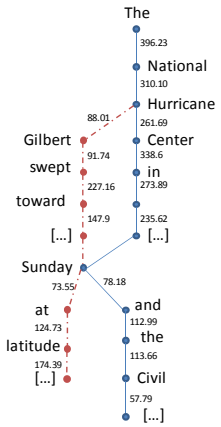


Figure 1. Example of a word graph representation

B. Obtaining New Sentences

In order to produce a new sentence, we employ Dijkstra's algorithm [24] to find the shortest paths between an initial node and the remaining ones that are directly or indirectly connected with it. We chose the shortest path algorithm because, on the one hand, it has been shown to be appropriate for compressing sentences in previous work [16], and on the other hand, the shortest path will also look for minimal-length sentences that contains information from several ones, thus allowing them to include more information.

Two strategies are proposed for defining a starting node to apply the searching algorithm over the document's graph representation:

- The initial node corresponds to the **first word in each sentence**. This manner, we ensure that for each sentence in the source document, we have at least one

derived sentence, so the whole content of the document is covered.
- The initial node corresponds to the **10 words with highest *tf-idf***. Term frequency-inverse document frequency (tf-idf) accounts for frequent terms in the document, but not very frequent in the whole collection of documents. With this strategy, we keep the most important terms and the information related to them.

C. Ensuring Sentences' Correctness

By applying the Dijkstra's algorithm over the graph we obtain all possible shortest paths between one node and the remaining ones. This leads to a high number of resulting sentences, which are not equally good. In fact, some of the sentences might be completely incomprehensible and not correctly formed. In order to guarantee the completeness and correctness of a new sentence, we define three basic constraints in order to discard those sentences, which do not satisfied all of them:

- The minimal length for a sentence must be 3 words (i.e., subject+verb+object).
- Every sentence must contain a verb.
- The sentence should not end in an article (e.g., a, the), a preposition (e.g., of), an interrogative word (e.g., who), nor a conjunction (e.g., and).

The remaining sentences after applying the aforementioned constraints will be used for building the abstractive summaries.

IV. Producing Abstractive Summaries

In order to use the new generated sentences (Section III) for building abstractive summaries, it is necessary to identify, which of them carry the most relevant information, since the length of the summary is restricted (in our case, to 100 words), thus not being possible to include all of them. Therefore, for determining important content we employed COMPENDIUM TS approach [25].

With the purpose of analysing whether or not it is better to generate new information before selecting the most important one, or the opposite (i.e., to extract relevant information first, and then generate new sentences from it), we apply COMPENDIUM in two different ways:

1) The set of new sentences obtained from the word graph-based method is the input for COMPENDIUM (***Graphs*+COMPENDIUM**).
2) The important content of the document is first selected, and then the word graph-based method is applied for generating new sentences derived from the extract (**COMPENDIUM+*Graphs***).

In both cases, the resulting summaries will be abstracts, since they do not reproduce verbatim the sentences of the source document.

V. Experimental Setup

In this section, we explain the dataset used, the experiments carried out as well as the results obtained together with an in-depth discussion.

A. Dataset

As dataset we randomly selected 50 documents of the DUC 2002 newswire corpus [26], each document having 500 words on average. Additionally, two model summaries written by humans are also provided for each document. These summaries have a length of approximately 100 words, which corresponds to a 20% compression rate with respect to the source documents.

B. Experiments and Results

In order to test the appropriateness of our suggested method for generating abstractive summaries, we followed the same guidelines as in DUC 2002 [27] (i.e., we produce generic single-document summaries of 100 words each) and we compare our abstractive summaries to the existing model summaries.

In addition to the two approaches explained in Section IV: ***Graphs*+COMPENDIUM** and **COMPENDIUM+*Graphs***, we define a ***baseline***, in which we generate the new sentences from the source document and select the first ones to build the abstractive summary, until the length of 100 words is reached.

Moreover, in order to broaden this analysis, we experiment with three heuristics concerning the length of the generated sentences:

- **ALL:** all generated sentences;
- **LONG:** only those sentences that are longer (in number of words) than the average length, and
- **SHORT:** only those sentences, which are shorter than the average length.

In total we analyze 18 types of abstracts: 2 strategies for generating new sentences, 3 summarization approaches, and 3 heuristics for selecting sentences with regard to their length, resulting in 900 different summaries (50 documents x 18 types).

For assessing the appropriateness of the generated abstractive summaries, we compare them to the model summaries employing the evaluation tool ROUGE [28]. In particular, we use the following metrics: ROUGE-1, ROUGE-2 and ROUGE-SU4, which account for the number of common unigrams, bigrams, and skip-bigrams with four words in-between at most, respectively. Tables I and II show the F-measure results for the abstractive summaries.

C. Discussion

As can be seen from both tables, results are not very high, though they are promising for further research, since they quantify how far we are from producing abstracts. They also help us to identify the limitations and the main challenges

 ISBN: 978-1-61208-162-5

Table I
RESULTS (F-MEASURE) OF THE ABSTRACTIVE SUMMARIES WHEN THE FIRST WORD OF EACH SENTENCE IS USED FOR GENERATING NEW SENTENCES.

Abstractive Approach	R-1	R-2	R-SU4
baseline-ALL	0.18726	0.04908	0.05967
baseline-LONG	0.19625	0.05029	0.06277
baseline-SHORT	0.20793	0.04877	0.06312
Graphs+COMPENDIUM-ALL	0.21609	0.05719	0.06951
Graphs+COMPENDIUM-LONG	0.22829	0.06187	0.07446
Graphs+COMPENDIUM-SHORT	0.21252	0.04808	0.06448
COMPENDIUM+Graphs-ALL	**0.29788**	**0.09663**	**0.11110**
COMPENDIUM+Graphs-LONG	0.29022	0.09660	0.10942
COMPENDIUM+Graphs-SHORT	0.16984	0.04633	0.05565

Table II
RESULTS (F-MEASURE) OF THE ABSTRACTIVE SUMMARIES WHEN THE TOP 10 WORDS WITH HIGHEST TF-IDF OF EACH SENTENCE ARE USED FOR GENERATING NEW SENTENCES.

Abstractive Approach	R-1	R-2	R-SU4
baseline-ALL	0.13058	0.03436	0.03957
baseline-LONG	0.14590	0.03729	0.04362
baseline-SHORT	0.15916	0.03604	0.04605
Graphs+COMPENDIUM-ALL	0.15668	0.04135	0.05042
Graphs+COMPENDIUM-LONG	0.17754	0.04554	0.05490
Graphs+COMPENDIUM-SHORT	0.17512	0.04228	0.05234
COMPENDIUM+Graphs-ALL	0.20850	0.06210	0.07048
COMPENDIUM+Graphs-LONG	**0.22323**	**0.06688**	**0.07647**
COMPENDIUM+Graphs-SHORT	0.18057	0.05186	0.05770

we need to face. A clear tendency is observed in the majority of the cases that the best results are obtained when the important information is first identified and extracted, and then the new sentences are generated (COMPENDIUM+*Graphs*). ROUGE results for COMPENDIUM+*Graphs-ALL* improve on average 55% with respect to the *Graphs*+COMPENDIUM-*ALL* approach when the first words of a sentence are used to generate the new sentences. The same approach but in the case of the top 10 words with highest tf-idf are used, leads to an improvement of 40% compared to the results obtained for *Graphs*+COMPENDIUM-*ALL*. Concerning the COMPENDIUM+*Graphs-ALL* and *baseline-ALL* approaches, the results of the former increase by 80% and 72%, for the first words or the top 10 words with highest tf-idf, respectively. In general, results are lower when the new sentences are generated from the words with highest tf-idf values. This is due to the fact that the summarization guidelines we followed together with the model summaries we had, focused on generic summarization, whereas our proposed strategy for generating new sentences using the top 10 words with highest tf-idf may be more appropriate to query-focused summarization, since this type of summary contains the most important information with regard to a specific topic, and consequently, the tf-idf method can provide some clues about the relevant topics of a document.

Now, by examining the content of the generated abstracts, we mainly focus on two types of problems. On the one hand, we try to elucidate the reasons why the *Graph*+COMPENDIUM approach performs worse than the COMPENDIUM+*Graph*, and on the other hand, we want to analyze the reasons of the low overall performance of the abstractive approaches.

Regarding the first type of analysis carried out, if we use the word graph-based method for generating new sentences first, and use all of them as input for COMPENDIUM, this TS tool can have difficulties in selecting important content. This occurs because many of the sentences will start with the same words (e.g., if we take the top 10 words with highest tf-idf), so once COMPENDIUM detects a specific fragment of information as relevant, sentences containing the same portion of information that have not been detected as redundant will be also selected, leading to summaries that have not much variation in content. In order to solve this limitation, besides checking for the correctness of the sentences once they have been generated and filtering out those ones, which do not satisfy the proposed constraints, we would also need to apply some constraints based on the information sentences contain, optimizing the set of generated sentences, so that only the best ones with respect to their content are used.

With respect to the general results of the abstractive approaches, since the length of the summaries is restricted to only 100 words, when selecting the most important sentences before or after generating new sentences, some of the concepts may not be included. Consequently, this affects the performance of the summaries, leading to low ROUGE results. Contrary to what was expected, longer sentences do not necessarily lead to better summaries, nor shorter sentences lead to more informative summaries. It happens the same problem as before: the concepts in the sentences may not present a great variation, focusing on a few topics, rather than providing an overview of the topics covered in the document. Finally, it is worth mentioning that producing pure abstracts is a challenging task, as it is shown also in previous research [18], where F-measure values for ROUGE-1 ranged from 13% to 18%.

VI. ADDRESSING THE LIMITATIONS OF THE APPROACH

The aim of this section is to conduct a preliminary analysis of the potential solutions to the problems previously identified.

A. Optimizing the Set of Generated Sentences

As it was previously stated, one possible solution for improving the selection the new generated sentences for taking part in the summary would be to find an optimization function that could provide us with the best generated sentences. In order to analyze if this could improve the final abstractive summaries, we carry out a preliminary experiment assuming an ideal case. We selected the 20% of the documents we used for our experiments, and we manually select the best sentences resulting from the word

 ISBN: 978-1-61208-162-5

graph-based method. As before, we used such sentences either as input for COMPENDIUM, or we first extracted the relevant content and then we generated the sentences, from which we manually selected the best ones. Table III shows the results of this pilot experiment. We perform a t-test to account for the significance of the results for a 95% confidence interval (results which are statistically significant are marked with a star).

Table III
ROUGE-1 RESULTS FOR THE ABSTRACTIVE SUMMARIES ASSUMING AN IDEAL CASE.

Abstractive Approach	Recall	Precision	$F_\beta = 1$
Graphs+COMPENDIUM_firstWords	0.207	0.209	0.208
Graphs+COMPENDIUM_firstWords_ideal	0.279*	0.287*	0.283
COMPENDIUM+Graphs_firstWords	0.283	0.291	0.287
COMPENDIUM+Graphs_firstWords_ideal	**0.293**	**0.301**	**0.296**
Graphs+COMPENDIUM_top10tfidf	0.197	0.199	0.198
Graphs+COMPENDIUM_top10tfidf_ideal	0.255*	0.263*	0.259*
COMPENDIUM+Graphs_top10tfidf	0.271	0.220	0.218
COMPENDIUM+Graphs_top10tfidf_ideal	0.283*	0.292*	0.287*

Assuming this ideal case, the results are improved by 25% on average, with respect to the original approaches. Furthermore, the improvement is higher for the *Graphs*+COMPENDIUM approach (36% and 31%, for rows 1-2 and 5-6, respectively). As it was previously shown, it is more appropriate to determine relevant information first by means of an extractive TS approach, and then try to compress and combine such information. In an ideal case, results for COMPENDIUM+*Graphs* improve by 5% and 10% with respect to *Graphs*+COMPENDIUM when the first words or the 10 words with highest tf-idf values are used for generating new sentences, respectively.

B. Combining Extractive and Abstractive Information

Here we want to analyze to what extent the generated sentences can be used in combination with extracts corresponding to the same documents. Therefore, we again experimented with the 20% of the documents and we took as a basis the extractive summaries for each of them generated by COMPENDIUM. Further on, taking also into consideration the abstractive summaries produced, we combined both types of summaries, according to these rules: i) if the sentence in the extract has one or more equivalent sentences in the abstract, we substitute the former for the latter; ii) if the sentence in the extract does not correspond to any sentence in the abstract, we keep the sentence in the extract, and iii) if the abstract contains some sentences that are not present in the extract, we enrich the extract with these sentences. In this manner, the new summary produced contains both extractive and abstractive information. Table IV shows the preliminary results of this experiment. Statistical differences according to a t-test are indicated with a star.

As it can be seen from the results obtained, we can confirm that for generic summaries, it is better to generate

Table IV
ROUGE-1 RESULTS FOR THE EXTRACTIVE+ABSTRACTIVE SUMMARIES.

Approach	Recall	Precision	$F_\beta = 1$
Extractive summary (ES)	**0.491**	0.456	0.472
ES+Graphs+COMPENDIUM_firstWords_ideal	0.471	**0.492**	**0.480***
ES+COMPENDIUM+Graphs_firstWords_ideal	0.426	0.458	0.441
ES+Graphs+COMPENDIUM_top10tfidf_ideal	0.458	0.456	0.457
ES+COMPENDIUM+Graphs_top10tfidf_ideal	0.405	0.436	0.419

the new sentences from the first words of each original sentence. Consequently, the summary will cover a wide range of topics. Moreover, results have improved considerably with respect to the ones obtained for the abstracts shown in Table III (62% on average). It is worth mentioning that when we take as a basis an extractive summary and we enrich it with abstractive information generated from the source document, F-measure results improve significantly compared to the initial extract. This is very positive result, since it indicates that we can carry out research into this type of summaries, improving the quality of them, as well as going beyond the simple selection of sentences.

VII. CONCLUSION AND FUTURE WORK

In this paper, we analyzed a method based on word graphs for generating abstractive summaries. The purpose of the method was to compress and merge information from sentences. In order to decide which of the new sentences should be included in the abstractive summary, we employed an extractive TS approach (i.e., COMPENDIUM), so that the most relevant sentences could be selected and extracted. We analyzed different strategies for generating abstracts, including the most appropriate way to generate new sentences, the order to select important information, and the length of the sentences. The results obtained, although encouraging, showed the difficulty of the task itself, and brought some insights of the problems with the resulting abstracts. In light of this, we conducted two additional preliminary experiments to analyze how to improve the resulting summaries. The main conclusion we can draw from this research is that the word graph-based method proposed is appropriate to generate abstractive information that can be later used to enrich extractive information, influencing positively in the resulting summaries.

Nevertheless, there is still a lot of room for improvement, so several actions have to be taken for further work. In the short-term, we plan to increase the corpus size and carry out the same experimentation with more documents, improving also the word-graph method. Moreover, we want to verify if the proposed strategy for generating new sentences taking into account the words with highest tf-idf could be appropriate for query-focused summarization. In the long-term, we are interested in analyzing other methods for representing information and how it can be generalized (e.g., concept

 ISBN: 978-1-61208-162-5

graphs).

ACKNOWLEDGMENT

This research has been supported by the FPI grant (BES-2007-16268) from the Spanish Ministry of Science and Innovation, under the project TEXT-MESS (TIN2006-15265-C06-01) and project grant no. TIN2009-13391-C04-01, both funded by the Spanish Government. It has been also funded by the Valencian Government (grant no. PROMETEO/2009/119 and ACOMP/2011/001).

REFERENCES

[1] K. Spärck Jones, "Automatic summarising: The state of the art," *Information Processing & Management*, vol. 43, no. 6, pp. 1449–1481, 2007.

[2] E. Hovy, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2005, ch. Text Summarization, pp. 583–598.

[3] D. S. Leite, L. H. M. Rino, T. A. S. Pardo, and M. d. G. V. Nunes, "Extractive automatic summarization: Does more linguistic knowledge make a difference?" in *Proceedings of the 2nd Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*. ACL, 2007, pp. 17–24.

[4] P. Lal and S. Rüger, "Extract-based summarization with simplification," in *Workshop on Text Summarization in conjunction with the ACL*, 2002.

[5] M. Liu, W. Li, and Q. Wu, M.ingli andLu, "Extractive summarization based on event term clustering," in *Proceedings of the 45th Annual Meeting of the ACL*, 2007, pp. 185–188.

[6] K.-F. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008, pp. 985–992.

[7] G. Carenini and J. C. K. Cheung, "Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality," in *Proceedings of the 5th International Natural Language Generation Conference, ACL*, 2008, pp. 33–40.

[8] C. Sauper and R. Barzilay, "Automatically generating wikipedia articles: A structure-aware approach," in *Proceedings of the 47th Association of Computational Linguistics*, 2009, pp. 208–216.

[9] H. Saggion, "Learning predicate insertion rules for document abstracting," in *Computational Linguistics and Intelligent Text Processing*, ser. LNCS, 2011, vol. 6609, pp. 301–312.

[10] ——, "A classification algorithm for predicting the structure of summaries," in *Proceedings of the Workshop on Language Generation and Summarisation*. ACL, 2009, pp. 31–38.

[11] D. Zajic, B. J. Dorr, J. Lin, and R. Schwartz, "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks," *Information Processing & Management*, vol. 43, no. 6, pp. 1549–1570, 2007.

[12] J. Clarke and M. Lapata, "Models for sentence compression: a comparison across domains, training requirements and evaluation measures," in *Proceedings of the 44th Annual Meeting of the ACL*, 2006, pp. 377–384.

[13] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," *Computational Linguistics*, vol. 31, pp. 297–328, 2005.

[14] E. Marsi and E. Krahmer, "Explorations in sentence fusion," in *Proceedings of the 10th European Workshop on Natural Language Generation*, 2005.

[15] K. Filippova and M. Strube, "Sentence fusion via dependency graph compression," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 177–185.

[16] K. Filippova, "Multi-sentence compression: Finding shortest paths in word graphs," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 322–330.

[17] A. Schrijver, *Theory of linear and integer programming*. John Wiley & Sons, Inc., 1986.

[18] F. Liu and Y. Liu, "From extractive to abstractive meeting summaries: can it be done by sentence compression?" in *Proceedings of the ACL-IJCNLP Conference Short Papers*. ACL, 2009, pp. 261–264.

[19] M. Galley and K. McKeown, "Lexicalized markov grammars for sentence compression," in *Proceedings of the Human Language Technology Conference of the NAACL*, 2007, pp. 180–187.

[20] J. Steinberger, M. Turchi, M. Kabadjov, R. Steinberger, and N. Cristianini, "Wrapping up a summary: From representation to generation," in *Proceedings of the ACL 2010 Conference*, 2010, pp. 382–386.

[21] Python-graph, "http://code.google.com/p/python-graph/," Last access: July, 2011.

[22] TreeTagger, "http://www.ims.uni-stuttgart.de/projekte/corplex/treetagger/," Last access: July, 2011.

[23] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer Networks ISDN Systems*, vol. 30, pp. 107–117, 1998.

[24] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[25] E. Lloret, "Text summarisation based on human language technologies and its applications," Ph.D. dissertation, University of Alicante, Spain, June 2011.

[26] DUC Past Data, "http://www-nlpir.nist.gov/projects/duc/data.html," Last access: July, 2011.

[27] DUC, "http://www-nlpir.nist.gov/projects/duc/guidelines.html," Last access: July, 2011.

[28] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Proceedings of ACL Text Summarization Workshop*, 2004, pp. 74–81.

 ISBN: 978-1-61208-162-5

Improving Email Management

Tõnu Tamme, Ulrich Norbisrath, Georg Singer, Eero Vainikko
Institute of Computer Science, University of Tartu
Tartu, Estonia
tonu.tamme@ut.ee, ulrich.norbisrath@ut.ee, georg.singer@ut.ee, eero.vainikko@ut.ee

***Abstract*—For twenty years, email has been the prominent means of computerized communication. Each day we receive a growing number of email messages from different origins, related to different topics, people, and locations. Some belong to the professional sphere, others are private. Usually we keep our messages in the inbox or store them in several mostly manually created hierarchical folders. Showing information in hierarchies and lists can be nowadays amended by views which allow a more explorative approach to access this. The goal of this paper is to analyze the automatic information management capabilities of present standard email clients and webmail services, show their shortcomings, and show some improvements of them through the use of auto categorization and graph exploration. We show that categorization is not supported by traditional email tools but that it facilitates discovery of new relations between email messages and therefore improves email management.**

***Keywords*—email; exploratory search; categorization; n-grams; ontology.**

I. INTRODUCTION

Email has undoubtedly conquered the position of being the most important means for written communication. Business and private life without email is unthinkable. However, this success has unavoidably entailed another challenge. The sheer amount of messages that we receive each day and the lack of appropriate tools even lowers productivity in some cases.

Today, we keep our messages in the main folder or store them in several specialized folders of our email program — one for each person or a group of persons, one for each institution, topic, or project. Most e-mail clients offer filters to do some of this sorting automatically, based on various keyword-like criteria. In spite of these sorting options, working through your emails and addressing relevant emails for your daily tasks becomes harder and harder with the growing load.

As a motivating example, we assume that we are working for a company, which is being audited. The auditor demands a list of ongoing projects between us and our clients and their interconnections. Therefore, we have to create a report describing each project, involved and related persons, and compile the material these persons worked on. We assume that there is no central file containing this information. The only sources are the email conversations between the employees and the respective clients (including various email attachments). The amount of emails to consider will be more than several thousands. In this paper, we will in particular look at the possibility to list all attachments between a group of persons and creating interconnections via conversation topics.

Addressing this with current email clients offers the following options: Standard email programs like Outlook, Thunderbird and webmail services like Gmail, Hotmail, or Yahoo mail allow manual sorting of email by sender, recipient, date received and date sent. Furthermore, they allow to use standard keyword search over the body and over some parts of the header. Gmail is the only solution offering a working and highly performing full text index. Gmail and Thunderbird both allow manual tagging, but with an increasing amount of tags, the management effort makes it less usable. All standard email systems support basic filtering functionality using patterns. In case of having emails sorted into person related folders, a task like described in the motivating example above will be time consuming and tedious as project related information is distributed over all those folders. If extensive manual tagging has been done in foresight of such a task, the effort will be significantly smaller. However, such extensive tagging is usually not done or not even supported by the email client.

To get an overview for such a report, exploring context related data in the emails would be very helpful. "Exploring" is used here in the sense of exploratory search. Exploratory search is defined in [1] like the following: *"Exploratory search can be used to describe an information-seeking problem context that is open-ended, persistent, and multi-faceted; and to describe information-seeking processes that are opportunistic, iterative, and multi-tactical. [...] In exploratory search people usually submit a tentative query to get them near relevant documents then explore the environment to better understand how to exploit it, selectively seeking and passively obtaining cues about where their next steps lie."* Exploratory search in emails is nowadays mainly based on browsing your own folders and tags as well as creating multiple queries for retrieving messages matching keywords.

There is more email search support from several recently emerged addons for Outlook and Gmail like Xobni [2], [3] or Xoopit [4]. Nevertheless, also these tools do not lighten the complexity in the aforementioned example. The shortcomings of the existing email solutions are mainly due to the following: None of the existing solutions have support for categorization via text analysis nor exploration apart from folder and tag browsing by any means. In this paper we suggest the combination of a categorization method and the automatic association of these detected categories with the mails and the involved persons.

This paper is organized as follows: Section II gives an overview of related work and tries to motivate our approach. Section III outlines our enhancement to classic email management. Section IV shows some results using these enhancements. Section V shows our conclusions and outlines future

 ISBN: 978-1-61208-162-5

work.

II. State of the art

This section will give an overview of the features of today's email systems and possibilities to analyze and enhance them. The focus is on their support of search and information management and their possibility to enhance automatic email management. We will analyze the methods and algorithms for automatic categorization and tagging on the one hand and their realization in email clients and email client plug-ins on the other hand. For a more practical and case driven analysis, we use the Enron Email Dataset. The dataset is a collection of 500.000 emails, organized in folders, that contains information from 150 senior management staff members at Enron. It is one of the few substantial "real" email repositories that was made available to the public for the purpose of improving email tools.

A. Email clients and email clients plug-ins

A study from June 2011 with over a billion of emails [5] shows a market share of email clients of Microsoft Outlook with 27% followed by 16% iOS devices (iPhone, iPad and iPod Touch), 12% Hotmail, 11% Apple Mail and 9% Yahoo Mail. Gmail has 7% market share being used as an email client, Android 1.7% and Thunderbird 1.2%. Microsoft Outlook, Apple Mail and iPhone Mail are the only clients, not coming with a full text search function over a full text index. Outlook can be enriched with a full text index through Google Desktop or Microsoft Search. Microsoft Outlook, GMail, and Thunderbird support manual tagging, but interoperability between different mail systems (with an exception of Gmail's IMAP tag emulation) is very limited. The share of the clients is recently shifting heavily to mobile clients (like iPhone mail and Android, which is basically Gmail).

Standard email clients involve different visualization and tagging techniques. For example Mozilla Thunderbird categorizes search results into suitable time intervals like years, months, or days of month.

One of our main complaints about most existing email clients is their lack of support for networking emails with each other. They only support manually sorting emails in hierarchic folders. Without duplication, it is not possible to assign an email to several topics. Tagging supports such a way of sorting, but usually breaks the order of hierarchies and therefore allows only flat ordering. Tagging allows to add special markers to emails with the aim of grouping similar messages together and making their finding easier without having to sort the mail in a strictly hierarchical directory structure. In some clients a different name might be used for this function, such as marking, labeling, categorizing, or adding keywords. Some clients (Mozilla Thunderbird, Microsoft Outlook, Opera Mail, Gmail) also enable the user to define new tags. Due to the flat nature of such tags, using tags does not really help to create structure. Therefore, there is usually only a limited set of tags suggested. For example Thunderbird and Gmail suggest personal and work tags. Gmail also has by default a separate tag for traveling purposes. Thunderbird and Opera also have a todo tag for emails indicating an assignment or task. Often it is possible to tag spam or junk messages. This function helps the client to avoid unwanted messages automatically in the future. Opera offers also "send reply", "call back" , "funny", and "valuable" tags by default. Apple Mail does not support tagging of emails, it has some support for flagging emails like staring mails in Gmail or marking mails as junk. It allows to create virtual folders, this means different views, but this does not allow to define tags as these are only views not folders where emails can be moved.

An interesting search task is tracking the attachments exchanged between two persons. The task is not trivial as it involves two persons and there must also be an opportunity to check the attachments. If such an attachment check is not part of the search, sorting can also be used. An attachment-based search is not possible in Yahoo Mail and Windows Live Mail because they lack the OR operator that enables matching the same name for both sender and receiver. The task is not difficult in Microsoft Outlook, but in Thunderbird and Opera Mail sorting has to be used to filter out messages with attachments.

There are several plug-ins available, enhancing the experience in email systems:

Xobni [2], [3] is the most common protagonist of the group of Outlook add-ons. As a plug-in it allows keyword based search and people based navigation in their Outlook mailboxes. Xobni extends the support for person related information and operates as an integrating platform between Outlook and online services like Facebook, Twitter, and LinkedIn. It automatically creates profiles of persons and their connections. These profiles contain statistics about relationships, contact information, threaded conversations, shared attachments, and information on that contact pulled from earlier mentioned online services.

Nelson Email Organizer (NEO) [6] is an Outlook add-on allowing different views of all messages in the inbox. It offers different views in different tabs and allows to organize messages in these tabs by date, by sender, or by attachment. All views support full text index keyword search. These are helpful features. However, they address no major conceptual simplification.

The Firefox extension *Xoopit* [4] turns Gmail into a robust, searchable media management tool for every piece of media that comes through the inbox. By indexing every attachment as well as every link to photos and videos from sites like Flickr, Picasa, and YouTube, Xoopit allows to easily search for and find any picture or video and view it from directly inside Gmail. XOOPIT was acquired by Yahoo in July 2009 and is now integrated into the Yahoo mail environment.

TaQuilla [7] is a Thunderbird extension which can be trained to "soft tag" incoming emails for different categories depending on an existing training set. It uses Bayesian analysis of the tags already given on stored emails to do this classification. In order to make the Bayesian filter work for a new tag, the existing messages have to be trained by showing the system examples of emails that carry the tag and messages that do not carry the specific tag. This needs to be done manually at the beginning for a certain amount of messages

 ISBN: 978-1-61208-162-5

until the system can take over. For example as outlined in [8] a user applied Taquilla to automatically separate "Personal" and "Business" related (not personal) messages. On the "Personal" side Taquilla analyzed a weekly general mailing from the user's church's pastor and came up with tokens like life, pastor, worship, sunday, thinking, and children. All those tokens showed an over 80% probability that the messages should be tagged as personal. On the business side on the other hand, it let Tacquilla analyze a posting on the Thunderbird testing list that was not tagged personal. It showed for other tokens like advance, feedback, bug, vista, and Thunderbird with an under 10% probability that these tokens were personal.

B. Integrative frameworks

ClearContext [9] is an email experience enhancing Outlook plugin. It has a very task driven integrational approach. It is very contact focused and supports scraping of appointment data from the emails. It helps filing, prioritizing emails based on sender, unsubscribing from conversation, defering emails to come back to your inbox, and various task and appointment management functions. It integrates these features into convenient workflows. Task management and communication is lightly integrated, but our idea of networking all the information of communication participants, topics, dates, and content into a free explorable graph is not realized.

Radar [10] is a research prototype also trying to advance task and email integration. It consists of several assistants employing machine learning. It supports automatic categorization of emails and scraping of relevant data out of emails and using this for guiding users through corresponding tasks. Regrettably, the product is not available as prototype or at the market.

Windows Search, *Spotlight* (Apple), *Google Desktop* are very similar. They all create a full text index of all files and emails stored on your computer and in case of Google also of mails stored on your Google account and make them accessible in a local Desktop query based keyword search. There is no special exploration support [11], [12], [13], [14], [15].

None of these tools or systems allow the exploration of the data pool contained in a standard email repository. Radar and ClearContext offer via their task integration some light support, but not in a graph-based manner. Thus, the interaction with today's email systems is still typically based on keyword search queries and browsing of manually maintained folder or tag hierarchies and the referenced messages.

As shown in the paper by Singer et. al. [16] keyword based search alone does not cover many of nowadays information needs. This is of course also true in the email context. The results support the hypothesis that also in the email context a graph based exploratory approach will significantly improve the search experience.

C. Automatic categorization and tagging

As already mentioned in the context of the example described above, manual tagging can be a tedious process. As machine learning algorithms become better with progresses in research, it makes sense to use these to facilitate the tagging or categorization process. Dredze et al. present an approach for generating summary keywords for emails in [17]. Their approach selects a set of keywords describing a single message in context to the topics of all messages. Their method is unsupervised. They are using Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDS) to determine a minimalistic amount of possible latent concepts. They use for their research the very versatile toolkit MALLET [18]. MALLET stands for MAchine Learning for LanguagE Toolkit. It is based on Java and supports statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning mechanisms applicable for text-based problems. Due to its huge variety of tools for text-processing, it is very well suited for carrying out experiments for topic and categorization analysis. It supports also the just mentioned LDS method.

It is also possible to use n-gram-based Text Categorization [19]. This method uses fingerprints of training texts and allows to compare other texts to these fingerprints. One of the main usages for this method is language detection. N-Gram distribution is very specific to languages, making it a very good choice for language detection. LibTextCat [20] is a freely available software supporting this method.

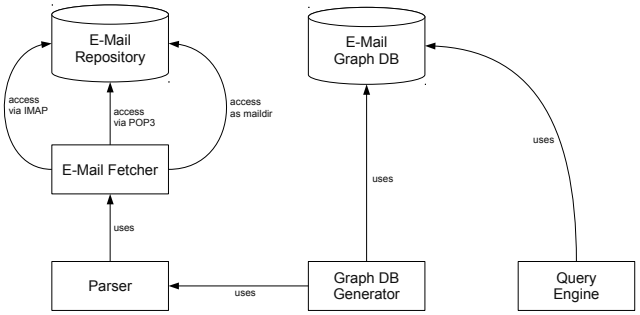


Figure 1. Email experimentation environment architecture.

III. PROPOSAL

We have shown in Section II that current email clients have problems with handling attachment search and creating interconnections via topics.

To overcome these shortcomings we have built an experimental email handling environment. The tool consists of four components: email fetcher, parser, database generator, and query engine. The architecture is depicted in Figure 1. The parts are acting autonomously. Thus they can be easily replaced with other suitable software for experimentation purposes.

The email fetcher reads the contents of an email account using the IMAP protocol. It facilitates retrieving information from different accounts. The email parser transforms a mailbox into a list of messages. Each email is a list of attribute value pairs. The body of a message is parsed into a three level list: paragraphs consist of lines and lines itself consist of words.

The first two components are written in Python as it has special libraries for dealing with emails and mailboxes.

 ISBN: 978-1-61208-162-5

Database generation and query engine are implemented with Prolog that has built-in support for presenting relations and exploring the search space due to its predicate structure.

The database generator transforms a list of email messages into a graph database that consists of labeled nodes and edges. For example a message node may have related nodes for its sender, receiver, subject, or date. A sender of a message may be receiver of another message. Thus, the graph database does not need to duplicate their nodes.

The text-based query interface provides a set of tools for visualizing and analyzing the message database. We can observe all the messages in a mailbox. We can also filter out all the emails from (or to) a person or between two persons. This can be done by tracing the values of From, To and Cc-attributes of messages.

In addition to using existent person→message relations Prolog helps in constructing new virtual relations on top of existent ones. To allow inferential relations between messages it will be beneficial to assign to each message 5–10 keywords. This provides us with alternative views to our email collection that can produce novel deduced relations not covered by the existing thread structure. With the help of message→topic relations we can build two new relations: person→topic and topic→person.

Several text analysis methods can be used to determine the topic of a message. Our first choice was to calculate the word frequency table. This method is language dependent and must be amended with a suitable list of stopwords and a stem picking routine.

The relative frequency method eliminates the need for using a stopword list. We calculate the frequency tables for the whole message set and for a chosen message. Comparing those frequencies for a word in the message we can choose potential keywords as words whose local frequency in a message is bigger than the global frequency in our whole text corpus. The frequency tables can be given in advance or built in the course of the analysis.

Some experiments to classify text with the help of n-grams have been carried out. This method relies on the availability of sample texts about potential topics.

For our classification we also link to large freely available ontologies like WordNet [21] and OpenCyc [22]. The ontologies consist of concepts, their definitions, and corresponding terms. Words acting as terms in an ontology are good candidates for a topic of an email message. Concepts and their hyperonyms are the basis of the conceptual network.

We have tried those methods to find the topics of a message automatically. Then we add the topics to the existing email graph and execute on it Prolog queries to resolve indirect connections between messages.

IV. Experiments

We made experiments with several mailboxes and chose as a testbed the Enron email dataset that has been made publicly available by the federal commission after the bankruptcy of the American company. The dataset consists of about 150 personal mailboxes and half million messages. A sample Enron message looks like the following:

```
Message no 844
From = John Arnold
To = Frank Hayden
Date = Fri, 14 Jul 2000 19:46:00 +0200
Subject = Re: Market Opinion about AGA's
Interesting observation...but I'm not sure I
    agree. I think consensus opinion is that
    anything under 2.7 TCF is very dangerous
    entering the winter. A month ago, analysts
     were predicting we would end the
    injection season with around 2.6 -2.7 in
    the ground. /.../
```

With the method described in the last section, we store it in our graph database in which the node numbers correspond to the internal construction of the graph. This will look like the following:

```
email(844).
node(844, 'XXX').
edge(844, 845, body).
node(845, [[['Interesting', 'observation...but
    ', "I'm", not, sure, 'I', 'agree.', 'I',
    think, consensus], [opinion, is, that,
    anything, under, '2.7', 'TCF', is, very,
    dangerous, entering, the], ['winter.', 'A
    ', month, 'ago,', analysts, were,
    predicting, we, would, end, the, injection
    ], [season, with, around, '2.6', -, '2.7',
     in, the, 'ground.', ...] ...]]).
edge(844, 847, subject).
node(847, 'Re: Market Opinion about AGA\'s').
edge(844, 848, from).
node(848, '"John Arnold"').
edge(844, 850, to).
node(850, '"Frank Hayden"').
edge(844, 851, date).
node(851, 'Fri, 14 Jul 2000 19:46:00 +0200').
```

As we have a graph database, a query for all attachments between groups of people is very easy to realize in contrast to classic email clients, which only rarely and insufficiently support such a query. In the following example, we list all the emails containing attachments exchanged between John Arnold and Mark Sagel. A person's name can be identified as a substring in his email-address. While collecting data the following basic query is also formatting the output.

```
?- F='"John␣Arnold"', T='"Mark␣Sagel"',
|      findall(X:From->To:Subject:Date->Attachment,(
|      email(X,Id,From,To,Date,Subject,Body),
|      (sub_atom(From,_,_,_,F), sub_atom(To,_,_,_,T)
|      ; sub_atom(From,_,_,_,T), sub_atom(To,_,_,_,F
    )),
|      edge(X, Y, 'attachment'), node(Y,Attachment))
    ,List),
|      print_list(List), length(List,N).
```

The output contains similar records as the query is not optimized to detect duplicate messages. The Enron dataset contains several of such duplicates due to its hierarchic nature.

```
3380:"John␣Arnold"->"Mark␣Sagel" <msagel@home.com> :
     Re: Natural gas update:Mon, 14 May 2001
    09:33:00 +0200->ng051301.doc
3389:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Natural gas update:Mon,
    14 May 2001 00:23:00 +0200->ng051301.doc
3397:"John␣Arnold"->"Mark␣Sagel" <msagel@home.com> :
     Re: Service Agreement:Mon, 27 Nov 2000 19:48:00
     +0200->Agree-Enron.doc
```

 ISBN: 978-1-61208-162-5

```
3406:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Service Agreement:Mon, 27
     Nov 2000 18:16:00 +0200->Agree-Enron.doc
3414:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Status:Thu, 16 Nov 2000
    19:30:00 +0200->energy2000-1112.doc
3422:"John␣Arnold"->"Mark␣Sagel" <msagel@home.com> :
     Re: Natural gas update:Mon, 14 May 2001
    09:33:00 +0200->ng051301.doc
3431:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Natural gas update:Mon,
    14 May 2001 00:23:00 +0200->ng051301.doc
3439:"John␣Arnold"->"Mark␣Sagel" <msagel@home.com> :
     Re: Service Agreement:Mon, 27 Nov 2000 19:48:00
     +0200->Agree-Enron.doc

...

3545:"John␣Arnold"->"Mark␣Sagel" <msagel@home.com> :
     Re: Service Agreement:Mon, 27 Nov 2000 19:48:00
     +0200->Agree-Enron.doc
3554:"John␣Arnold"->"Mark␣Sagel" <msagel@home.com> :
     Re: Natural gas update:Mon, 14 May 2001
    19:33:00 +0200->ng051301.doc
3572:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Natural gas update:Mon,
    14 May 2001 00:23:00 +0200->ng051301.doc
3580:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Service Agreement:Mon, 27
     Nov 2000 18:16:00 +0200->Agree-Enron.doc
3588:"Mark␣Sagel" <msagel@home.com> -> "John␣Arnold"
     <jarnold@enron.com> : Status:Thu, 16 Nov 2000
    19:30:00 +0200->energy2000-1112.doc
F = '"John␣Arnold"',
T = '"Mark␣Sagel"',
List = [ (3380:'"John␣Arnold"'->'"Mark␣Sagel"␣<
    msagel@home.com>':'Re:␣Natural␣gas␣update':'Mon,
    ␣14␣May␣2001␣09:33:00␣+0200'->'ng051301.doc'),
    (3389:'"Mark␣Sagel"␣<msagel@home.com>'->'"John␣
    Arnold"␣<jarnold@enron.com>':'Natural␣gas␣update
    ':'Mon,␣14␣May␣2001␣00:23:00␣+0200'->'ng051301.
    doc'), (3397:'"John␣Arnold"'->'"Mark␣Sagel"␣<
    msagel@home.com>':'Re:␣Service␣Agreement':'Mon,␣
    27␣Nov␣2000␣19:48:00␣+0200'->'Agree-Enron.doc'),
     (3406:'"Mark␣Sagel"␣<msagel@home.com>'->'"John␣
    Arnold"␣<jarnold@enron.com>':'Service␣Agreement'
    :'Mon,␣27␣Nov␣2000␣18:16:00␣+0200'->'Agree-Enron
    .doc'), (3414:'"Mark␣Sagel"␣<msagel@home.com>'->
    '"John␣Arnold"␣<jarnold@enron.com>':'Status':'
    Thu,␣16␣Nov␣2000␣19:30:00␣+0200'->'energy2000
    -1112.doc'), (3422:'"John␣Arnold"'->'"Mark␣Sagel
    "␣<msagel@home.com>':... : ... -> 'ng051301.doc'
    ), (3431:'"Mark␣Sagel"␣<msagel@home.com>'->... :
     ... -> 'ng051301.doc'), (... : ... -> ... ->
    ...), (... -> ...)|...],
N = 21.
```

The query can be enhanced and stored as a predefined predicate. The sorted list without duplicates looks like the following (Prolog code for actual sorting and removing duplicates omitted):

```
?- attachments_between('"John␣Arnold"','"Mark␣Sagel"
    ').
16 Nov 2000 19:30 ("Mark␣Sagel"  -> "John␣Arnold")
    Status          (energy2000-1112.doc)
27 Nov 2000 18:16 ("Mark␣Sagel"  -> "John␣Arnold")
    Service Agreeme (Agree-Enron.doc    )
27 Nov 2000 19:48 ("John␣Arnold" -> "Mark␣Sagel" )
    Re: Service Agr (Agree-Enron.doc    )
14 May 2001 00:23 ("Mark␣Sagel"  -> "John␣Arnold")
    Natural gas upd (ng051301.doc       )
14 May 2001 09:33 ("John␣Arnold" -> "Mark␣Sagel" )
    Re: Natural gas (ng051301.doc       )
14 May 2001 19:33 ("John␣Arnold" -> "Mark␣Sagel" )
    Re: Natural gas (ng051301.doc       )
true.
```

This result shows how beneficial a graph based structure is for creating a query about the exchange of attachments.

We have amended our emails via topic detection mechanisms discussed in the state of the art section with related keywords. Looking at the initial email example from John Arnold to Frank Hayden, we observe that John Arnold is concerned about events in winter. Thus we can try to find links to other messages dealing with this season. To construct the keyword list we first filter out stopwords and other ill-formed strings of messages like MIME code and HTML tags. We discover that the word "winter" is listed among the top ten keywords of the following message:

```
?- keywords_message(1128).
--Top10
5, prompt
5, $
3, position
3, futures
2, winter
2, stress
2, spread
2, scenarios
2, payout
2, normal
```

This message has a different subject than our initial message ("Re: Stress Testing"). Therefore, we have discovered here an interconnection between two different threads, solving our second described problem of finding such interconnections through exploration. We are able to discover two persons who are linked via a topic they discuss but who are not directly related via a thread. This is a property not derivable in a classic email system setup. The underlying automatically generated graph structure makes it possible to explore this relation.

V. Conclusions and Future Work

In this paper, we have given an overview of the state of the art of automatic information management capabilities of today's email clients and webmail services. We have discussed their shortcomings and have presented ideas for their improvement by using auto categorization and graph exploration. Email has been the prominent means of computerized communication for a quarter of a century. However the core technology has remained unchanged. Especially an update of the information management capabilities seems to be needed for dealing with an ever increasing number of emails. We also have presented the results of our experiments with our email handling tool.

In our future work it will be possible to extend the capabilities of the email handling tool by providing a simpler user interface and extending it with graph navigation and visualization facilities. We are also planning to switch from n-gram and frequency based categorization to LDS based categorization and apply our methodology to a bigger dataset.

Acknowledgment

This paper was supported by the European Social Fund through the Estonian Doctoral School in Information and Communication Technology.

 ISBN: 978-1-61208-162-5

REFERENCES

[1] R. W. White, G. Marchionini, and G. Muresan. Evaluating exploratory search systems: Introduction to special topic issue of information processing and management. *Information Processing and Management*, 44:433–436, 2008.

[2] D. E. Descy. Microsoft add-ons and updates. *TechTrends*, 54(2):7–8, 2010.

[3] R. L. Scott. Wired to the world: Xobni. *North Carolina Libraries*, 66(3):64, 2009.

[4] Yahoo! acquires xoopit. Available from: http://www.myphotos.yahoo.com/ [cited August 24, 2010].

[5] Email client popularity: June 2011. Available from: http://www.campaignmonitor.com/stats/email-clients/ [cited July 17, 2011].

[6] NEO – the microsoft outlook email software Add-On. Available from: http://www.caelo.com/ [cited August 24, 2010].

[7] R. Kent James. MesQuilla » TaQuilla. Available from: http://mesquilla.com/extensions/taquilla/ [cited August 18, 2010].

[8] R. Kent James. MesQuilla » blog archive » TaQuilla provides automatic "soft" tags for messages. Available from: http://mesquilla.com/2009/02/26/taquilla-provides-automatic-soft-tags-for-messages/ [cited March 7, 2011].

[9] ClearContext – Outlook plugin to organize email and manage inbox. Available from: http://www.clearcontext.com/ [cited August 18, 2010].

[10] M. Freed, J. Carbonell, G. Gordon, J. Hayes, B. A. Myers, D. Siewiorek, S. Smith, A. Steinfeld, and A. Tomasic. Radar: A personal assistant that learns to reduce email overload. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1287–1293, 2008.

[11] P. A. Chirita, R. Gavriloaie, S. Ghita, W. Nejdl, and R. Paiu. Activity based metadata for semantic desktop search. *The Semantic Web: Research and Applications*, pages 439–454, 2005.

[12] D. Aumüller and S. Auer. Towards a semantic wiki experience–desktop integration and interactivity in WikSAR. In *Semantic Desktop Workshop*, 2005.

[13] Duen Horng Chau, Brad Myers, and Andrew Faulring. What to do when search fails: finding information by association. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 999–1008, Florence, Italy, 2008. ACM.

[14] D. Pogue. Google takes on your desktop. *New York Times*, 2004.

[15] B. Turnbull, B. Blundell, and J. Slay. Google desktop as a source of digital evidence. *International Journal of Digital Evidence*, 5(1):1–12, 2006.

[16] Georg Singer, Ulrich Norbisrath, Eero Vainikko, Hannu Kikkas, and Dirk Lewandowski. Search-Logger – analyzing exploratory search tasks. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 751–756. ACM, 2011.

[17] M. Dredze, H. M. Wallach, D. Puller, and F. Pereira. Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 199–206, 2008.

[18] MALLET homepage. Available from: http://mallet.cs.umass.edu/ [cited March 7, 2011].

[19] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[20] libTextCat – lightweight text categorization. Available from: http://software.wise-guys.nl/libtextcat/ [cited August 20, 2010].

[21] WordNet: A lexical database for English. Available from: http://wordnet.princeton.edu/ [cited July 17, 2011].

[22] Cycorp, Inc. Available from: http://www.cyc.com/ [cited July 17, 2011].

 ISBN: 978-1-61208-162-5

A Redundant Bi-Dimensional Indexing Scheme for Three-Dimensional Trajectories

Antonio d'Acierno
Institute of Food Science
National Research Council
Avellino, Italy
dacierno.a@isa.cnr.it

Alessia Saggese, Mario Vento
DIEII
Universitá degli Studi di Salerno
Fisciano (SA), Italy
{*asaggese,mvento*}*@unisa.it*

***Abstract*—The need of efficient methods for querying continuously moving object databases arises in many applications of intelligent video surveillance. As a consequence, several data indexing strategies have been introduced in order to improve data storing and retrieving and develop more efficient trajectory analysis systems. However, even though efficient spatial indexes in bi-dimensional planes are usually available, several issues occur when data to be handled are three- or even four-dimensional as, for instance, moving objects trajectories in real world environments. For this reason, we are interested in proposing a new indexing scheme capable of analysing and retrieving three-dimensional trajectories in efficient way. This goal is achieved by redundantly projecting and analysing a collection of trajectories on bi-dimensional planes and validating the obtained result through a clipping algorithm. Experimental results show that the proposed approach yields good performance in terms of averaged retrieving time when applied to time interval queries.**

Keywords-MOD; Three-dimensional trajectory; Indexing; Time interval query.

I. INTRODUCTION

Moving Object Databases (MODs) are used to store continuously moving objects. According to the widely adopted line segments model [1], the object motion is expressed through its trajectory; trajectories, in turn, are represented by a polyline in a three-dimensional space, the first two dimensions being referred to space and the third one to time (Figure 1).

The demand of efficiently querying MODs arises in many contexts, from air traffic control to mobile communication systems. There are at least two categories of queries that are worth to be considered: queries about the future positions of objects, and queries about the historical positions of moving objects. Historical queries can be further classified [1] into coordinate-based queries and trajectory based queries. While trajectory-based queries involve information about a trajectory such as topology and velocity, coordinate-based queries in turn include:

1) Time interval queries: select all objects within a given area and within a given time period;
2) Time-slice queries: select all the objects present in a given area at a given time instant;
3) Nearest neighbor queries: select the k nearest neighbor objects to a given point in space at a given time instant.

A key problem to be addressed concerns the indexing of these data. R-trees, proposed by Guttman in his pioneering paper [2], was a widely adopted solution motivated by the Very Large Scale Integration (VLSI) design: how to efficiently answer whether an area is already covered by a chip. R-trees hierarchically organize geometric objects representing them using the MBRs (*Minimum Bounding Rectangles*); each internal node corresponds to the MBR that bounds its children while a leave contains pointers to objects. Starting from the original structure, several optimizations have been proposed [3]; in [1], for example, the particularities of spatio-temporal data are captured by two access methods (STR-tree and TB-tree) while SEB-trees [4] segment space and time.

When the aim is to index and query repositories of large trajectories, the size of MBRs can be reduced segmenting each trajectory and then indexing each sub-trajectory using R-Trees; such an approach is described, for example, in [5], where a dynamic programming algorithm to minimize the I/O for an average size query is proposed. SETI [6] segments trajectories and groups sub-trajectories into a collection of *spatial partitions*; queries run over the partitions that are most relevant for the query itself. TrajStore [7] co-locates on a disk block (or in a collection of near blocks) trajectory segments using an adaptive multi-level grid; thanks to this method, it is possible to answer a query only reading a few blocks.

Our main aim is to extend a video surveillance system [8] with an efficient method for querying continuously moving object databases in order to interpret the behaviour of different entities populating a scene. Even though efficient bi-dimensional indexing methods are usually available, several problems arise when data to be handled are three- or even four-dimensional as happens for the considered video surveillance system. Indeed, this framework identifies a real object by using a triple (x, y, f) where (x, y) represents the

 ISBN: 978-1-61208-162-5

object position whereas f is the frame number; assuming a constant frame rate, the frame number and the time can be used as synonyms. In order to achieve this aim, we extend the existing video surveillance system by proposing a new indexing scheme capable of analysing and retrieving three-dimensional trajectories in efficient way. The proposed method works by redundantly projecting and analysing a collection of trajectories on bi-dimensional planes. Obtained result is finally validated in the three-dimensional plane through a clipping algorithm. Different experiments, performed by using the POSTGIS [9] system, will show that the proposed approach yields good performance in terms of averaged retrieving time when applied to time-interval queries on synthetic data.

II. The Proposed Solution

A trajectory is usually referred to as a list of space-time points:

$$< (x_1, y_1, t_1), (x_2, y_2, t_2), ..., (x_N, y_N, t_N) >$$

where the generic pair (x_i, y_i) is the spatial location and t_i represents the time. Each point is thus treated as an object in an extended spatial domain, since time is considered as an additional dimension. As already mentioned, we use the line segments model [1], each segment being the linear interpolant between two consecutive points.

To answer a time-interval query, we have to verify the intersection between a 3D query box, identified by bottom-left-back $(x_{min}, y_{min}, t_{min})$ and top-right-front $(x_{max}, y_{max}, t_{max})$ points, and all the segments of each trajectory. To determine if a line segment lies inside, outside or partially outside the box, we can use a clipping algorithm; one of the most efficient methods for our purposes is the extension to 3D of the 2D Cohen-Sutherland Line Clipping Algorithm [10].

The recursive bi-dimensional Cohen-Sutherland Line Clipping Algorithm considers only segment endpoints; if at least one endpoint of the segment s lies inside the clip box, the hypothesis *h: s intersects the box* can be trivially accepted. If both endpoints are outside the clip box, the segment may or may not intersect with the clip box. In some cases *h* can be still trivially accepted (as it happens for segments AB in Figure 2) or rejected (segment CD). Other situations (segment EF and segment GH) can be solved recursively by subdividing the line into two segments and using the extensions of the clip box edge; one of the obtained segment can be trivially rejected, while the other one is the new segment to be analyzed.

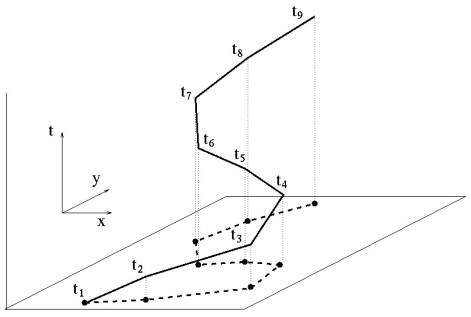


Figure 1. An example of spatio-temporal trajectory; *x* and *y* dimensions refer to position while the third dimension (*t*) refers to time.

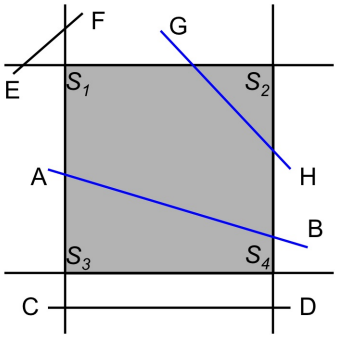


Figure 2. Some segments lying inside (AB and GH) or outside (CD and EF) the clipping area $S_1S_2S_3S_4$.

The bi-dimensional Cohen-Sutherland algorithm can be easily extended to the 3D case [10]; here, operations have to be performed with reference to six half-planes ($y < y_{min}$, $y > y_{max}$, $x < x_{min}$, $x > x_{max}$, $t < t_{min}$, $t > t_{max}$) and by considering the obtained twenty-seven regions.

In the worst case, when the trajectory does not intersect the box, we have to verify all the segments in the trajectory; such an approach is too expensive for a large amount of trajectory data, thus the aim of the proposed indexing strategy is to reduce the number of candidate trajectories to clipping, taking advantage of the existing 2D indexes.

The method we propose is based on three derived bi-dimensional spaces obtained by projecting each 3D trajectory onto (X, Y), (X, T) and (Y, T) planes. It is worth to observe that if a trajectory intersects the 3D query box, then each trajectory projection will intersect the correspondent query box projection. This is a necessary but not sufficient condition since the opposite is clearly not true: if all projections trajectory intersect correspondent box projection on considered spaces, they do not have to intersect the 3D query box too. To better explain this concept, Figure 3 shows a trajectory on 3D space and its projections on 2D spaces: we can note that all the trajectory projections intersect correspondent box projection, although the trajectory does not intersect the 3D query box.

Figure 4 resumes the main phases of the method needed to answer a time-interval query. For each three-dimensional trajectory t (Figure 4a), we redundantly store three bi-dimensional trajectories. Each trajectory is obtained by projecting t on the XY plane (t_{XY}), on the XT plane (t_{XT})

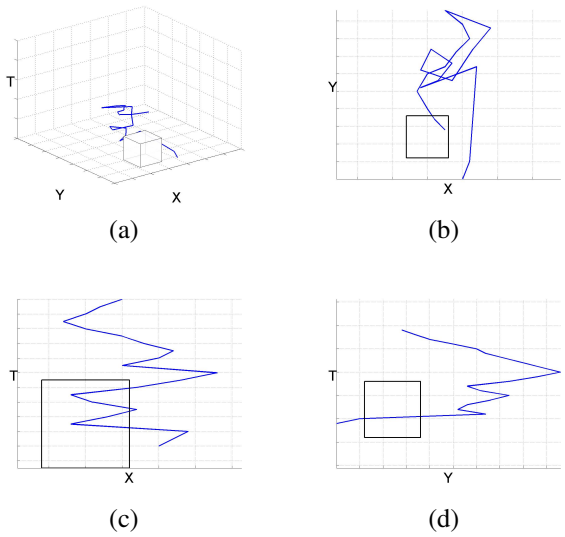


Figure 3. An example of trajectory (a) and its projections on the different coordinate planes XY (b), XT (c) and YT (d). Note that although the trajectory does not intersect the query box, its projections do it.

and on the YT plane (t_{YT}), as shown in Figure 4b. Given a box B representing the time-interval query to be solved, we similarly consider B_{XY}, B_{XT} and B_{YT}.

Using one of the available bi-dimensional indexes, we can find on each plane the following three trajectory sets in a very simple and efficient manner (Figure 4c):

$$T_{XY} = \{t_{XY} : MBR(t_{XY}) \cap B_{XY} \neq \emptyset\} \quad (1)$$

$$T_{XT} = \{t_{XT} : MBR(t_{XT}) \cap B_{XT} \neq \emptyset\} \quad (2)$$

$$T_{YT} = \{t_{YT} : MBR(t_{YT}) \cap B_{YT} \neq \emptyset\} \quad (3)$$

The set T of the candidate trajectories to be clipped in 3D space is thus trivially defined as:

$$T = \{t : t_{XY} \in T_{XY} \wedge t_{XT} \in T_{XT} \wedge t_{YT} \in T_{YT}\} \quad (4)$$

As shown in Figure 4d, the candidate set is composed by trajectories whose MBR on each plane intersects the corresponding projection of the query box; this does not imply that, for example, $t^i_{XY} \in T_{XY}$ actually intersects B_{XY}. This choice will be motivated in the last Section.

III. Experimental Results

We test our system over synthetic data sets generated as follows.

Let W and H be the width and the height of our scene; let S be the time interval we are interested in. Each trajectory starting point is randomly chosen in our scene at a random time instant t_1; the trajectory length is assumed to follow a Gaussian distribution. We also randomly chose an initial direction along the x axis (d_x) as well as a direction along the y axis (d_y).

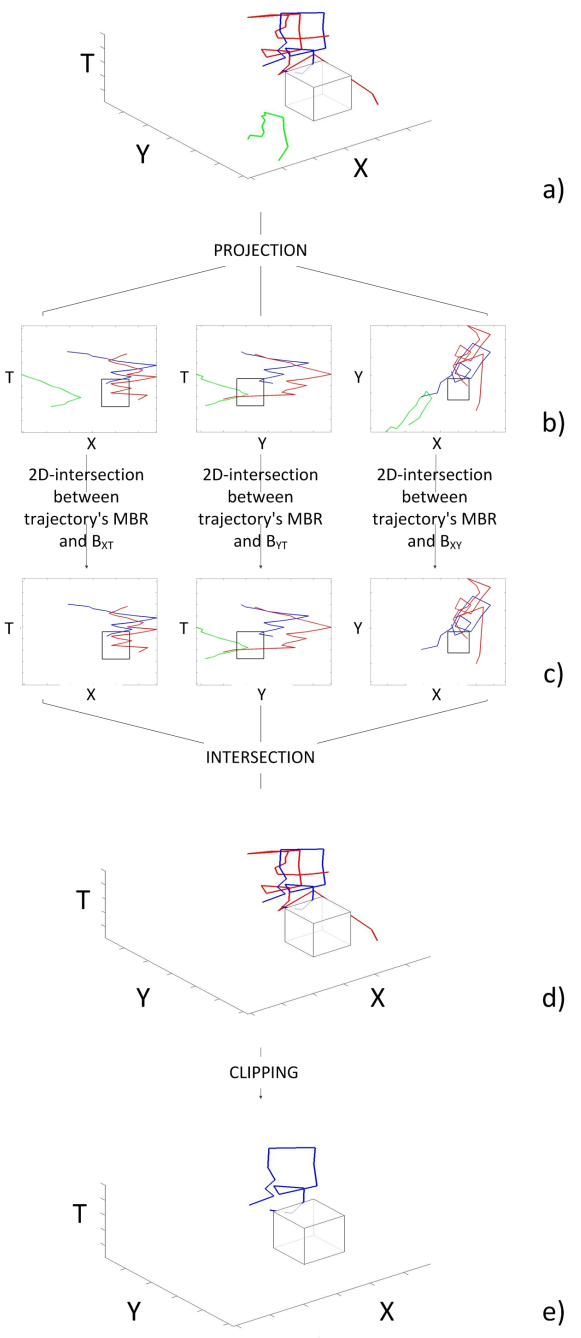


Figure 4. An overview of the proposed method. Figure a) shows a query box and some examples of trajectories; in b) there are the projections of each trajectory on the coordinate planes. Figure c) only shows the projections that intersect the correspondent query box. Figure d) shows the trajectories whose three projections intersect correspondent boxes. Finally, Figure e) shows the final result of our method, i.e., trajectories that really intersect the query box.

Table I
THE PARAMETERS USED TO GENERATE OUR DATA.

Scene width	W	10^3
Scene height	X	10^3
Time interval length in seconds	S	10^4
Number of trajectories (thousands)	T	{1,2,3,5,10}
Mean number of points in each trajectory (thousands)	$\overline{L}$	{1,2,3,4,5,10}
Standard deviation for trajectory's length	σ_L	10
Probability to invert the direction along x	PI_x	5%
Probability to invert the direction along y	PI_y	5%
Maximum velocity along x (pixels/seconds)	V_x^{max}	10
Maximum velocity along y (pixels/seconds)	V_y^{max}	10

At each time step t, we first generate the new direction, assuming that d_x (d_y) can be changed with probability PI_x (PI_y) and then we randomly chose the velocity along x and y (expressed in pixels/seconds and assumed to be greater than 0 and less than two fixed maxima). The new position of the object can be so easily evaluated; if it does not belong to our scene, d_x and/or d_y are changed.

We define the scene populated with trajectories as the "Scenario". Table I reports the free parameters to be chosen together with the values we chose to create the 30 different scenarios used in our experiments.

We store data in Postgres using PostGIS, an extension built to store and query spatial data like points, lines and polygons. We represent mobile object trajectory as a tuple of $(mId, mTraj_{XY}, mTraj_{XT}, mTraj_{YT})$, where mId is the unique trajectory identifier and $mTraj_{XY}$ (respectively $mTraj_{XT}$ and $mTraj_{YT}$) is the XY projection of the trajectory (respectively, the XT and the YT projections), represented as a sequence of segments (a PostGIS multi-line). Data are indexed using the R-tree over GIST (Generalized Search Trees) indexes [11] since it guarantees, compared with the PostGIS implementation of R-trees, best performances for spatial queries. Similarly to R-trees, GIST indexes break up data into a search tree according to their spatial position. Once data have been indexed, PostGIS provides a very efficient function to perform intersection between boxes and MBRs in a 2D space. It is clear that this kind of intersection could be not accurate, especially for large trajectories whose MBR, especially in the XY plane, could cover almost the entire area.

To test our system we need some queries, each query being defined by the corresponding three dimensional cube. The dimension D_c and the position P_c of the cube of course affect the obtained performance. We decided thus to test different dimensions, expressed as a percentage of the whole volume; we choose $D_c \in \{1\%, 5\%, 10\%, 20\%, 30\%, 50\%\}$. Each query is repeated several times (to be more precise, smaller cubes are queried more times) and results are then averaged.

The overall averaged querying time ($\overline{QT}$) of course depends on T, on $\overline{L}$ and on D_c; $\overline{QT}$ can be expressed as the sum of two terms: $\overline{QT_q}$ is the time needed to extract data from the database while $\overline{QT_c}$ is the time needed to apply the clipping algorithm to candidate trajectories.

We conduct our experiments on a PC equipped with an Intel quad core CPU at 2.66 GHz, using the 32 bit version of the PostrgreSQL 9.0 server and the 1.5 version of PostGIS. We obtain that, on average, $\frac{\overline{QT_c}}{\overline{QT_c}+\overline{QT_q}} = 50.4\%$. In the following we do not further investigate on $\overline{QT_c}$ and $\overline{QT_q}$ but we will concentrate on how $\overline{QT}$ increases as the free parameters vary.

Diamonds in Figure 5 express (in a log-log scale for the sake of readability) $\overline{QT}$ in seconds as the number of trajectories varies for different values of $\overline{L}$, both for small cubes ($D_c = 1\%$) and for large ones ($D_c = 30\%$). To analyze the relationship between $\overline{QT}$ and T we polynomially approximate $QT(T)$, both for each fixed D_c and for each $\overline{L}$. We obtain, with a very good approximation, that $\overline{QT}$ linearly increases with T (lines in Figure 5).

Diamonds in Figure 6 express $\overline{QT}$ in seconds as the dimensions of the cubes vary, having $\overline{L}$ as parameter and for several values of T. In this case we obtain that $\overline{QT}$ quadratically depends on D_c (lines in Figure 6).

Last diamonds in Figure 7 express $\overline{QT}$ in seconds as $\overline{L}$ varies for several values of D_c; in this case we have again a quadratic dependency on $\overline{L}$.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

In the framework of a video surveillance system, we are interested in efficiently querying a three dimensional MOD using off the shelf solutions and so, in this paper, we propose a redundant storing system to index large repositories of three dimensional trajectories using widely available two dimensional indexes; the proposed method has been implemented using PostGIS, the well known spatial extension of the PostgreSQL server. Preliminary results, obtained on time interval queries performed against synthetic data, show that the proposed solution is able to fully exploit retrieving capabilities based on well established two dimensional indexes.

Concerning our work in progress, it must be observed that there are several possibilities to improve the performance of our system. First, we have a querying time that linearly increases with T while is a quadratic function of $\overline{L}$; thus, it can be pointed out that it is better to have more trajectories

 ISBN: 978-1-61208-162-5

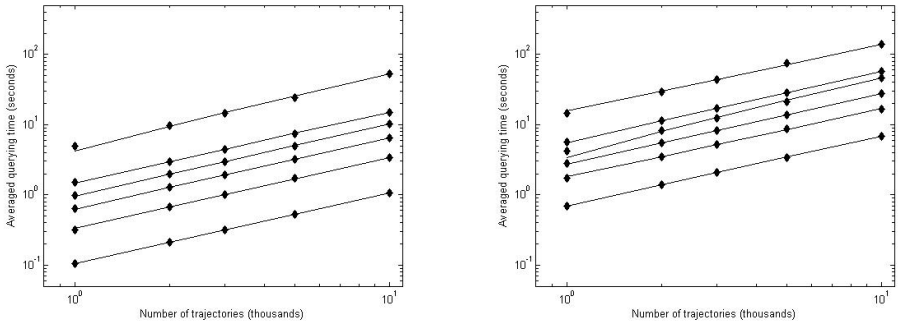


Figure 5. $\overline{QT}$ (in seconds) as the number of trajectories increases having the number of points in each trajectory (in thousands) as parameter.

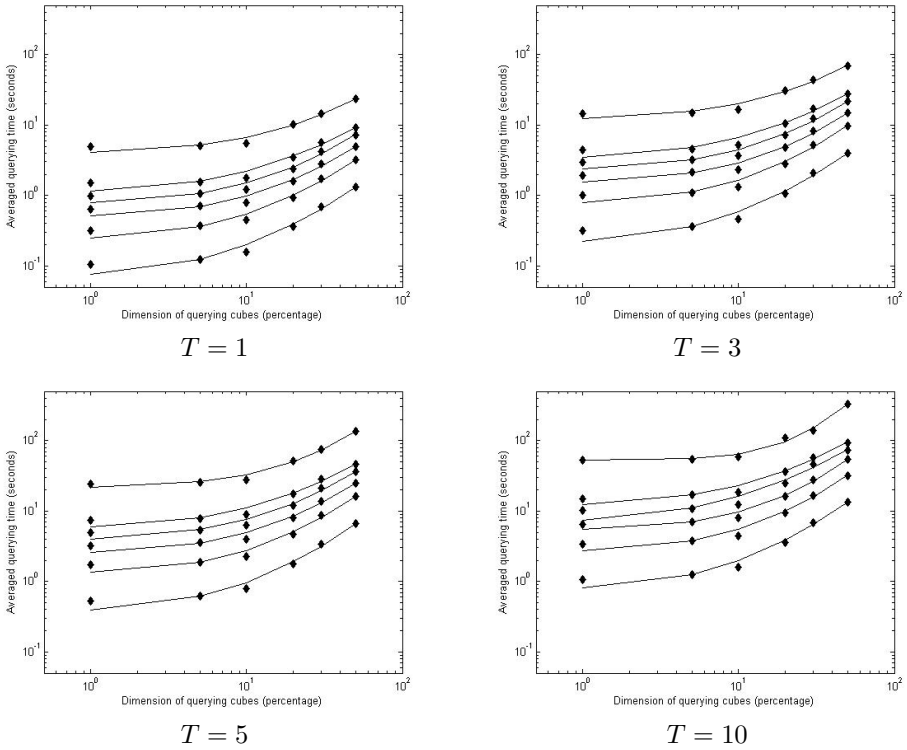


Figure 6. $\overline{QT}$ (in seconds) as the dimension of the querying cube (in percentage of the whole volume) increases and having $\overline{L}$ as parameter.

with fewer line segments and this can be obtained pursuing at least two strategies. A trajectory can be easily *compressed* because, in many context, data are highly redundant when sampling objects' positions at high rate. A trajectory can be then splitted in two or more sub-trajectories and applying our method to the set of sub-trajectories; such an approach, while clearly improves the performance due to the linear dependency of $\overline{QT}$ on T, also optimizes MBR-based indexes.

It is then worth to be noted that the clipping algorithm has to be applied in parallel to each candidate trajectory. This step can be easily implemented using multi threading, in order to take advantage from multi-core and multi-processors systems. On the other hand, the functions testing if two geometries intersect, available in GIS systems, are typically applied sequentially on each considered pair. For such a reason we query our DB (on each projected plane) for trajectories whose MBR intersects the corresponding projection of the query box (a very fast query given the MBR based index), instead of trajectories that actually intersect the corresponding MBR, that have to be tested sequentially.

Furthermore, we store data redundantly since, for each trajectory t, we store t_{XY}, t_{XT} and t_{YT} so that our

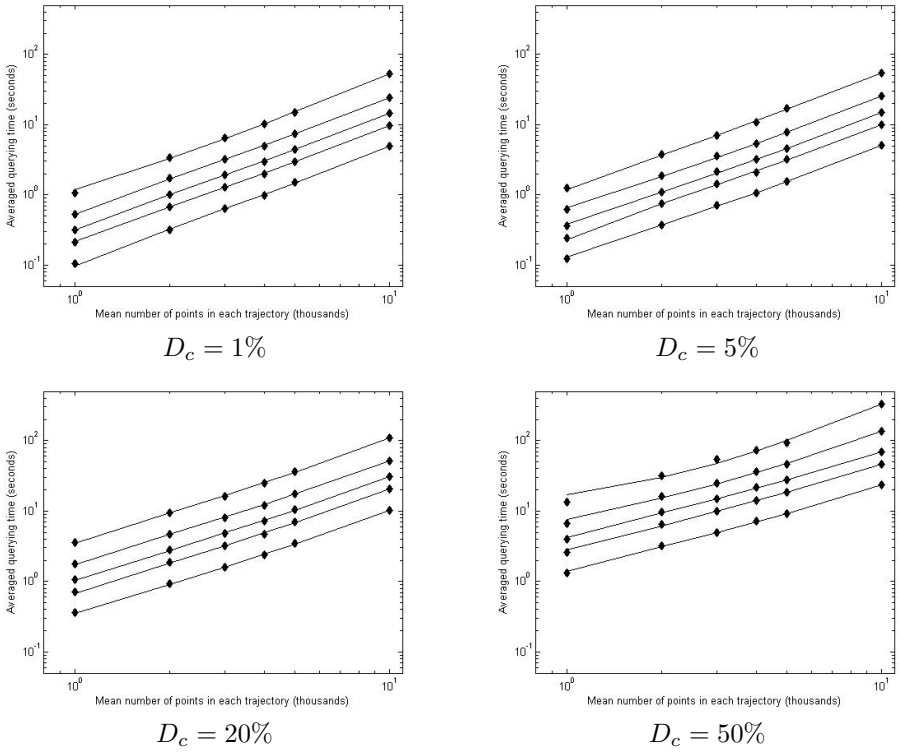


Figure 7. $\overline{QT}$ (in seconds) as the number of points in each trajectory (in thousands) increases and having the number of trajectories as parameter.

schema roughly doubles the used memory (for each three dimensional point we store three bi-dimensional points) and this can easily become a serious limitation when the number of stored trajectories increases. To overcome such a problem we are developing a new schema that heavily diminishes data redundancy.

The system then needs to be extended; in fact queries different from the time-interval ones are likely to be easily solvable with our solution.

Last, another objective is to make our system able to store and handle data as they are acquired.

REFERENCES

[1] D. Pfoser, C. S. Jensen, and Y. Theodoridis, "Novel approaches in query processing for moving object trajectories," in *Proceedings of VLDB Conference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 395–406.

[2] 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.

[3] Y. Manolopoulos, A. Nanopoulos, A. N. Papadopoulos, and Y. Theodoridis, *R-Trees: Theory and Applications*. Springer, 2005.

[4] Z. Song and N. Roussopoulos, "Seb-tree: An approach to index continuously moving objects," in *Proceedings of the 4th Conference on MDM*. London, UK, UK: Springer-Verlag, 2003.

[5] S. Rasetic, J. Sander, J. Elding, and M. A. Nascimento, "A trajectory splitting model for efficient spatio-temporal indexing," in *Proceedings of the 31st international Conference on VLDB*. VLDB Endowment, 2005, pp. 934–945.

[6] V. P. Chakka, A. Everspaugh, and J. M. Patel, "Indexing large trajectory data sets with seti," in *CIDR*, 2003.

[7] P. Cudre-Mauroux, E. Wu, and S. Madden, "Trajstore: An adaptive storage system for very large trajectory data sets," *Data Engineering, International Conference on*, vol. 0, pp. 109–120, 2010.

[8] D. Conte, P. Foggia, G. Percannella, and M. Vento, "Performance evaluation of a people tracking system on pets2009 database," in *Proceedings of the 7th IEEE International Conference on AVSS*, 2010, pp. 119–126.

[9] "POSTGIS," `http://postgis.refractions.net//`.

[10] 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.

[11] J. M. Hellerstein, J. F. Naughton, and J. F. Naughton, "Generalized search trees for database systems," in *Proceedings of the 21st VLDB Conference*. Zurich, Switzerland: Morgan Kaufmann Publishers Inc., 1995.

 ISBN: 978-1-61208-162-5

Reasoning on High Performance Computing Resources

An Urgent Computing Scenario

Axel Tenschert, Pierre Gilet
Service Management & Business Processes (SANE)
HLRS - High Performance Computing Center Stuttgart
70569 Stuttgart, Germany
E-mail: {tenschert, gilet}@hlrs.de

***Abstract* — The emergent growing amount of available information and data in the last decade has led to very large data stocks that express specific knowledge. This knowledge can be stored in ontologies. Reasoning strategies are then required to deal with a huge amount of data even if the allotted time frame to perform this task is restricted. This work covers the research issue of performing a time-wise restricted reasoning by means of ontologies. The presented approach is suitable for processing a reasoning strategy on high performance computing resources by considering a short time window and offering a solution for a quick allocation of required resources.**

Keywords - Ontology Matching; Reasoning; High Performance Computing; Resource Allocation

I. INTRODUCTION

The work described in this paper presents two research topics that need to be considered for solving the challenge of reasoning in a High Performance Computing (HPC) environment. Both research topics are explained in the next sections.

When thinking of an end user that performs a reasoning task with ontologies that are adequate for his/her needs it has to be taken into account that this end user might not be an expert in HPC infrastructures. This leads to the challenge of supporting the end user in a user friendly and time saving way in order to deal with the fact that time is short. At the present time, the allocation of computing resources at HLRS [1] is performed with a high level of effort and many human interactions involving an IT expert at HLRS that has to perform lots of manual steps. This manual workflow of computing resource allocation is time consuming. And if the end user is not an expert in HPC infrastructures the manual workflow is slowed further down. It is to be noted that HLRS is a federal high performance computing centre providing access to its HPC resources to researchers in Germany and Europe [2] and is currently extending its IT infrastructure through the acquisition of a new Cray XE6 platform in 2011 [3]. The move to production of that new supercomputer will increase the customer base of HLRS, which could comprise members of the traditional HPC community but also of non-HPC communities having only very few knowledge of HPC infrastructures. Hence, the need to support end users belonging to a non HPC community such as the semantic community is growing. To this end, the manual workflow solution is not an adequate option anymore.

Furthermore, a reasoning strategy is required that allows ontology matching over HPC resources. According to Ramesh and Gnanasekaran [4], the overall goal of an ontology matching is a merge of ontologies in order to create a new single terminology that can be used for reasoning purposes. The merge is an organization and reusage of concepts found in the source ontologies. This approach is used within this work to perform a matching between a set of ontologies in order to develop a resulting merged ontology. The resulting merged ontology is actually an ontology initially selected among the ontologies given as input that has been assigned the highest level of priority and is therefore called the priority ontology in the next sections. The merge process enriches that priority ontology with additional concepts, thereby leading to the generation of the resulting merged ontology. Also, the matching strategy considers similarities between the matched terms of the ontologies thanks to the use of a similarity value. This approach is also promoted by Pirró and Euzenat [5]. In their work, they describe a whole framework for the use of similarities in semantics.

The aim of this paper is thus to present an approach for designing a workflow going from allocating HPC resources up to performing a reasoning task in a merged ontology by means of the allocated computing resources. This publication is based on the ongoing PhD thesis written by Mr. Tenschert, and details about implementations and a result validation will be presented in future publications.

This paper gives an introduction to the described research field and related problems (I.), presents current research activities and challenges (II.), demonstrates a related use case scenario (III.), proposes a novel workflow for the given problem (IV.) and finally concludes with an outlook at future developments (V.).

II. CURRENT RESEARCH ACTIVITIES AND CHALLENGES

Nowadays, strategies for computing resource allocation and reservation in the HPC domain as well as reasoning strategies are available. However, the requirements and constraints imposed within an HPC environment are quite complex and specific to concrete use case scenarios, and reasoning strategies often need to deal with high amounts of

 ISBN: 978-1-61208-162-5

data also in a scenario where time restriction plays a prominent role.

The reservation and allocation of computing resources has been a research topic for HPC environments as well as for issues dealing with SLA (service level agreement) management and SLA lifecycles. The Grid Resource Allocation Agreement Protocol Working Group (GRAAP-WG) [15] offers solutions addressing this issue via the development of the Web Service Agreement specification (WS-Agreement) [7] allowing the creation of SLAs defining guarantee and service terms for the allocation of resources between two parties such as a service provider and a consumer. However, the use of HPC resources gives rise to the challenge of having to deal with very detailed and specific guarantee and service terms in an SLA. This results from the management of a specific IT infrastructure requiring very precise knowledge about the computing resources. This issue brings about the question of how to create a very specific SLA with specifications relating to HPC.

When thinking about reasoning strategies out of an information data pool - regardless whether the data is stored as text documents, graphics or visualized models - integration of information is of interest. The management of information by means of integration techniques is described by Lembo et al. [9] with a general formula (1).

$$<G, S, M> \quad (1)$$

- G is a global schema expressed in the global language L_G with the alphabet A_G. L_G determines the expressiveness allowed for specifying G;
- S is a set of local schemas modeled in the source language L_S with the alphabet A_S. A_S determines the set of defined constraints. A_S is disjoint from A_G;
- M is the mapping of G and S.

The presented formula for data integration is used for information retrieval approaches consisting in receiving targeted data and enhancing afterwards a source data set with the new available data. Such an approach is presented by Su et al. [10] for identifying an ontology matching strategy that makes use of a target ontology and a source ontology in order to enhance the source ontology.

Further, vector based techniques for dealing with word ambiguity are relevant for reasoning approaches in order to ensure the correct use of a term by validating its meaning. For instance, noteworthy vector based techniques are the latent semantic analysis (LSA) described by Landauer et al. [11] and random indexing described by Karlgren et al. [12], Chatterjee et al. [13] or Sahlgren [14].

Regarding reasoning, one can consider the research field of ontology matching as relevant, too. In this context, Zhang [16] describes the advantages of using ontologies and the semantic web for representing information. Gal et al. [17] discuss the need for ontology matching using matching of concepts with the aim to describe the meaning of data by considering heterogeneous distributed data sources and by also considering uncertainties in ontologies. Additionally, Huang et al. [18] give an overview of the use of ontologies relating to bioinformatics as formal knowledge representation models in order to offer knowledge to an expert. To this end, relevant ontology matching strategies are required for providing an expert with knowledge represented in the ontologies belonging to a various set of ontologies.

Due to the fact that various strategies for ontology matching have become more and more elaborate in the recent years, ontology matching approaches will be considered in this work as well.

III. URGENT REASONING SCENARIO

The use case scenario demonstrating the use of the resource allocation and reasoning workflow (RAAR-WF) is divided into two main parts:

1. Allocation of HPC resources,
2. Reasoning with a priority ontology.

In this scenario, an end user is in the need of receiving information split among various ontologies and the time frame for it is restricted. This means the results of the reasoning process have a validity window. Upon exceeding the time limit, the information becomes not required anymore and therefore has reached its point of validity. For instance, one could consider a biomedical scientist as an end user in need of receiving information (Fig. 1) for a patient or an urgent study before the point of validity is exceeded.

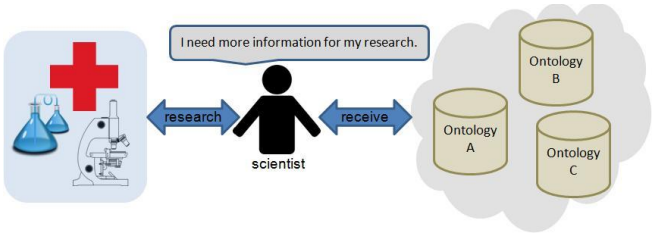


Figure 1: End User from the biomedical Research Area

The following subsections cover respectively the aforementioned two points (allocation of resources, and then reasoning) to ensure a good understanding of the whole scenario. The bringing together of both sections make up the RAAR-WF solution.

A. Allocation of HPC Resources

At present, the allocation of resources in an HPC environment (e.g., at HLRS) requires expert knowledge about IT infrastructures on the part of the end user. For instance, the end user has to know about the different architecture types, the characteristics of compute nodes, the installed software packages and tools, etc. Additionally, he/she must determine what configuration of computing resources best addresses the needs of the use case scenario. The challenge of allocating computing resources has increased due to the fact that within the HPC domain, Grid and Cloud infrastructures have become more and more complex. Considering a non expert end user in need of HPC, Grid or Cloud infrastructures, one can assume that the configuration of the requested computing resources will lead to a high level of effort for the end user and as well for the IT

 ISBN: 978-1-61208-162-5

expert who supports the end user. Besides, this process will be a time consuming one, too.

One can also note that the IT expert who supports the end user during the request for computing resources is also the person in charge for resource allocation. The IT expert performs a validity check of the end user request and then performs the resource allocation and reservation of the computing resources. The more requests are submitted to the IT expert, the more effort and time is obviously required to handle them. Timewise, in the urgent reasoning scenario, there is a high risk of exceeding the point of validity regarding the needed resources because of the general amount of requests for computing resources and the effort needed for supporting an end user who is not an expert in HPC, Grid or Cloud infrastructures.

B. Reasoning with a Priority Ontology

The foundation underlying the urgent reasoning scenario described in this work is a set of ontologies that fit the end users needs. In that scenario, the end user makes a selection of ontologies that most address the scenario requirements, e.g., ontologies about proteins or treatment modality.

One priority ontology needs to be defined first to make a clear distinction between the target and the source ontologies. One source ontology becomes the priority ontology and is then enhanced with the selection of target ontologies. Hence, the aim of the end user in this scenario is to improve one source ontology selected as the priority ontology in order to perform in the end a reasoning task with it. To this end, the end user can only choose one data source, the priority ontology, which will receive the required information. This strategy speeds up the overall reasoning procedure compared to a reasoning strategy that would force the end user to perform a reasoning task on the whole set of identified ontologies appropriate for the use case scenario.

The end user is provided with an automated merge of the selected ontologies that reduces the time for the end user to access the required information. However, one must ensure that the automated matching strategy produces information usable for the end user. To this purpose, a validity check during the matching process is required.

IV. IMPROVEMENTS THROUGH THE RAAR-WF

A. Improving the Allocation of HPC Resources

Thanks to the plugIT [6] approach, the process of validation and allocation of computing resources executed by the IT expert is enhanced in a way that the IT expert has only to approve the recommendations from the plugIT IT Socket. The general idea about the plugIT IT Socket is to support an end user who plays the role of a project applicant by means of the Online Proposal Submission (OPS) application. The OPS application supplies a form that the project applicant has to fill in to request access to computing resources. The project applicant thereby provides all the information required to run an automated assumption process that finds out the best HPC, Grid or Cloud configuration required for the scenario.

After receiving the recommendation from the plugIT IT Socket, the IT expert, who plays the role of the project approver, validates the recommendation and sends a notification with the recommendation back to the project applicant. The recommendation from the plugIT IT Socket is actually an SLA. For plugIT, the schema used as the foundation for the definition of the SLA XML structure is the WS-Agreement specification. However, in order to deal with the specific requirements relating to the HPC domain, additional elements were necessary. Those additional pieces were provided by the so-called WS-Agreement schema for HPC (HPC-WSAG [8]), which extends the WS-Agreement specification with HPC specific items. The recommendation proposed to the project approver is therefore an SLA offer based on the HPC-WSAG schema.

The plugIT IT Socket also requires information about the HPC site and its available infrastructure to make an assumption about the most fitting resource configuration to offer based on the input given by the project applicant. For this purpose, the IT infrastructure, the SLAs and special criteria for the SLAs are all represented as graphical models in an online repository accessible by the plugIT IT Socket. These models are designed by an IT expert, the infrastructure modeler, that has extensive knowledge about the existing HPC, Grid or Cloud environment and useful SLAs applicable to the available computing resources. The benefit of this approach is that the models are easily created by the infrastructure modeler when new computing resources are made available or if the current hardware changes. Thanks to this, the knowledge about the computing infrastructure can be shared among many IT experts. It is mapped in graphical models stored in the online repository. By means of this online repository, the plugIT IT Socket has enough information helping it produce SLA offers automatically.

The ability of the plugIT IT Socket to find out SLA offers based both on models stored in the online repository and on the input of the project applicant comes from the semantic kernel component of the plugIT IT Socket. The semantic kernel transforms the input of the project applicant into an ontology, and then compares it with the models representing the SLAs, SLA criteria and IT infrastructure that have been also transformed into model ontologies (MOs). Further, a domain ontology for HPC environments is used to support the transformation into MOs and the comparison between ontologies. This means that the functionality of the semantic kernel is twofold:

1. Transformation of the project applicant's input and models into ontologies and MOs,
2. Comparison of ontologies.

The picture of the resource allocation by means of the plugIT IT Socket is presented in Fig. 2.

 ISBN: 978-1-61208-162-5

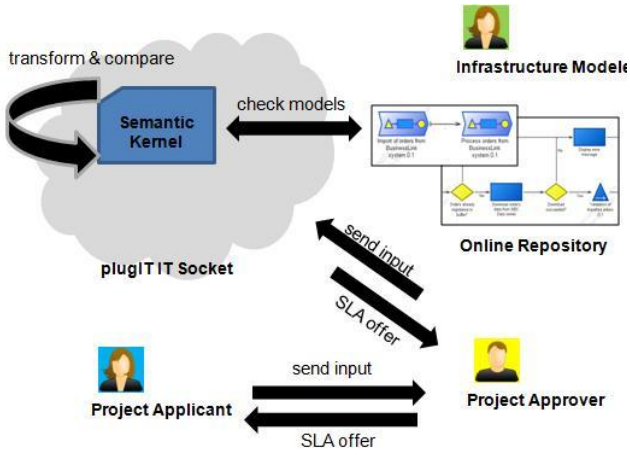


Figure 2: Resource Allocation Overview

The recommendation of SLAs to the project applicant via the plugIT IT Socket automates the process of resource allocation and makes it more efficient compared to the traditional manual workflow for resource allocation of HPC resources.

Additionally, concepts for the emerging cloud computing technology have been considered because of the fact that this new method for usage of distributed computing resources requires as well a clear strategy for resource allocation. The use of a cloud based approach provides the possibility to allocate needed computing resources for large scale data sets within a cloud testbed. One of the goals of the BonFIRE project is to develop a multi-site Cloud prototype. Within this scope, a cloud testbed is set up for research activities in the framework of the EU founded BonFIRE project [19]. The testbed will allow a large scale testing of research activities. This will be beneficial to the described work, especially considering the matching of ontologies made of very large data sets and matching procedure requiring vast amounts of computing power. The ongoing research and the results of this project, especially regarding the use of the cloud testbed, will influence the effective resource allocation as well.

B. Improving the Reasoning with a Priority Ontology

The use of a priority ontology enables reasoning on only one ontology that contains all the relevant information from a previously performed selection of ontologies. However, a strategy for matching the ontologies of the input set and further performing a validity check of the matching results is elaborated in this work with the aim to develop an ontology matching application. To perform adequate matching, a similarity value is created that expresses the level of accordance between matched entities. The matching workflow is performed in two major steps, the preparation and the execution, that are in turn subdivided.

- Preparation of similarity matching:
  a. Identification of relevant ontologies,
  b. Selection of relevant entities,
  c. Definition of the search space.
- Execution of similarity matching:
  a. Generation of the similarity value;
  b. Interpretation of the generated similarity value.

During the preparation phase, adequate ontologies are identified by the end user in order to create the set of target ontologies to be compared with the priority ontology. For this step the expertise of the end user is needed to decide which ontology to select. Then, the selection of the entities takes place to prepare the matching of the priority ontology entities to those of the target ontologies. For each entity of the priority ontology one matching iteration is performed. This means that the amount of matching iterations grows significantly with every entity of the priority ontology. However, the selection of the entities is based on the end user's expertise. This brings an additional possibility to specify the matching process in a very detailed fashion. The next step is the definition of the search space that establishes the number of neighboring entities that need to be taken into account for the matching of an entity. It is a necessary step in order to match the relations between entities. Due to the fact that the relation of one entity to its neighboring entities might not be similar in different ontologies, it is quite important to define the depth level, i.e., the search space, needed to assess the relations between entities in different ontologies. The deeper the search space is defined, the more numerous matching processes are performed and the higher the cost becomes which is associated with the matching process.

In the next step, the execution phase covers first the generation of the similarity value defining the level of compliance between entities. The similarity value is created out of a set of different values generated by various similarity matching processes using parameters such as the features of the concepts and relations to the neighboring concepts. The number of considered neighboring concepts was defined previously in the search space definition step. The second step of the execution phase is the interpretation of the similarity value. The similarity value is used for merging entities into the priority ontology based on the expressed level of compliance of the matched entities. Therefore, a high similarity value leads to a high probability of similarity between entities. Nevertheless, a validity check of the matching results is still required.

C. Improving the Reasoning with a Validity Check

Vector based techniques provide a solution for comparing the matching results saved in the priority ontology with the contents of a text document related to the topic of the use case scenario. The text document is provided by the end user having expertise in the considered topic. The selected text is the validity check document (VCD) containing text about the topic of the use case relating to the contents of the priority ontology. The selected terms from the priority ontology are concept and feature names. These terms are compared with those found in the VCD. Through random indexing techniques, the occurrence of a term coming from the VCD is calculated to determine how frequent it is found in the priority ontology. The random indexing approach for performing the validity check is divided into phases.

 ISBN: 978-1-61208-162-5

The random indexing solution is based on word space approaches and is therefore applicable to the required validity check with the VCD. Using word spaces means creating a high dimensional vector space for words to further construct a statistical value used for the next vector space. In the urgent reasoning scenario, the words considered for the validity check are the terms generated by the matching approach. These terms are concept and feature names. Regarding the vector space constructed by means of the previous statistical value, this strategy works with the assumption that if a set of words continuously appears in a text within the same context, then the meaning of the words will remain the same. It becomes thus possible to validate the terms of the priority ontology with those of the VCD and check if they can even be found in the VCD. However, in order to make an assumption about the context of the terms, the validity check examines the related terms in the VCD and in the priority ontology as well with the aim to compare those related terms. The analysis of the term relations in the VCD is done by examining the occurrence of words in the same sentence while the same analysis for the priority ontology is done by examining the relations between the concepts. A matrix containing the occurrence of terms in the priority ontology and in the VCD can be then created. This matrix is used to validate if the relations in the priority ontology appear in the VCD as well. Nevertheless, word space approaches face the challenge of scalability and efficiency. The use of HPC resources addresses this challenge, but a more fine grained approach for dealing with this issue in order to reduce the amount of required computing resources is still recommended. To this end, the simple vector based word space approach is enhanced through the use of the vector based random indexing approach that creates models such as those produced by latent semantic analysis (LSA) approaches. Following this approach, an extensive co-occurrence matrix is created first and then a reduction of the co-occurrence matrix is performed which limits the size of said matrix. Within the reduction phase, vectors of terms put in a specific context that occur multiple times are aggregated to accumulated context vectors. This way, the random indexing approach reduces the amount of required computing resources to perform the validity check in the given period of time.

Still, vector based techniques run the risk of producing unusable results when the text documents for comparison do not fit well the specific scenario needs. This leads to the question of whether the use of a vector based approach for a validity check as described previously is really usable without requiring a high level of effort from an expert who needs to do a very precise selection of text documents fitting the use case scenario. At the time of writing this paper, the vector based approaches are evaluated as well as the strategy of performing a validity check by means of a comparison ontology. The comparison ontology is matched with the priority ontology with the aim to make a proof of confidence regarding the updated priority ontology.

D. The RAAR-WF

The RAAR-WF comprises the allocation of best fitting HPC resources and the reasoning with the priority ontology including the validity check. The whole process going from allocating the required computing resource to getting back the urgently needed information through reasoning on one priority ontology is performed by considering a restricted time frame. The RAAR-WF is represented in Fig.3. As shown there, three points of human intervention can be identified in the workflow during the following phases:

1. Request for computing resources,
2. Configuration of the matching process and definition of the needed ontology set,
3. Reasoning with the created priority ontology.

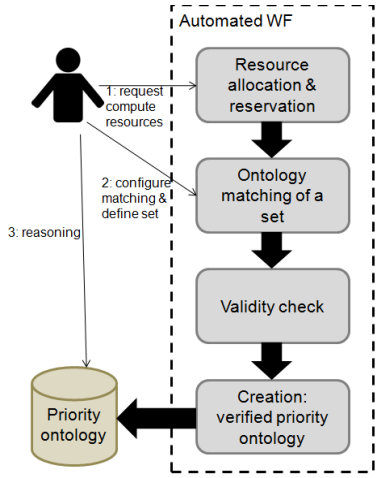


Figure 3: RAAR-WF

Beside the intervention of the end user, the most demanding tasks in terms of effort remain those performed in the automated part of the workflow. This includes the following steps:

1. Resource allocation and reservation: thanks to the use of the plugIT IT Socket, the resource allocation and reservation are provided in an automated fashion by the OPS application;
2. Ontology matching of a set: the selection of ontologies for the creation of the set is made by the end user, however the matching of the ontologies of the set is done automatically;
3. Validity check: the validity check guarantees the quality of the matching results thanks to an additional comparison of the matching results;
4. Creation of the verified ontology: after the validity check, the priority ontology is finally created based on the matching results and with the validity check taken into account.

The last step of the whole RAAR-WF workflow is the reasoning via the use of the priority ontology. The time constraint caused by the urgent computing scenario is also considered through the use of HPC, Grid or Cloud resources and a highly effective ontology matching method and

 ISBN: 978-1-61208-162-5

validation of matching results aiming at creating a priority ontology for reasoning.

V. CONCLUSIONS AND OUTLOOK

The described RAAR-WF includes a smart solution for automated HPC resource allocation involving graphical modelling and semantic processing that transforms models into ontologies and then compares the generated ontologies. The only steps to be taken over by human beings for the resource allocation and reservation are the creation of the necessary models and SLAs. The creation or update of models is only required if changes in the computing infrastructure have been made, such as the acquisition of a new cluster. The modelling effort is predictable and quite easy to make. It becomes thus possible to allocate and reserve HPC, Grid or Cloud resources within a short period of time. This complies with urgent computing cases such as the urgent reasoning scenario.

Furthermore, the reasoning strategy outlined in this document is performed on reserved computing resources which provide adequate computational power, and it makes use of a highly efficient ontology matching approach. The splitting of the ontology matching approach into a preparation phase and an execution phase offers a reliable matching solution whose output is checked by the validity check in order to guarantee reliable matching results. Also, the use of similarities increases further the reliability of the matching results. Since the result of the complete RAAR-WF is one single priority ontology, the task of the end user regarding the reasoning part is simplified because he/she has to consider only one ontology instead of having to cross check a set of many ontologies.

Regarding future developments, the validity check of the matching results offers the opportunity for further research dealing with the evaluation of various vector based word space approaches and diverse ontology matching strategies. The outcome of the work aiming at finding out which strategy has the highest probability of producing reliable matching results depends on the selected strategy as well as on the specific requirements and configurations of the use case scenario. Furthermore, the already mentioned BonFIRE project provides a cloud testbed for research activities. Therefore, the results obtained by that project will influence the resource allocation of computing resources in the HPC domain for this work. In the future, a cloud-like environment will be considered to allocate computing resources for the proposed ontology matching strategy.

ACKNOWLEDGMENT

This work has been supported by the plugIT project [6] and has been partly funded by the European Commission's ICT activity of the 7th Framework Programme under contract number 231430. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

The BonFIRE project has received research funding from the European Commission under the Information Communication Technologies Programme (ICT), contract number 257386. The project has a consortium of more than 13 partners from industry and academia as well as non-profit organizations.

REFERENCES

[1] High Performance Computing Center Stuttgart (HLRS), web site: http://www.hlrs.de/ (last accessed: July 22, 2011)

[2] HPC Europa, web site: http://www.hpc-europa.org/ (last access: July 22, 2011)

[3] Cray Wins Supercomputer Contract From the University of Stuttgart Valued at More Than $60 Million, http://investors.cray.com/phoenix.zhtml?c=98390&p=irol-newsArticle&ID=1486975&highlight (last accessed: July 22, 2011)

[4] C. Ramesh and A. Gnanasekaran, "Methodology Based Survey on Ontology Management", International Journal of Computer Sciences & Engineering Survey (IJCSES), vol. 1, no. 1, 2010

[5] G. Pirró and J. Euzenat, "A Semantic Similarity Framework Exploiting Multiple Parts-of Speech", Proceedings OTM, INRIA Grenoble Rhône-Alpes & LIG, 2010

[6] plugIT project, web site: http://plug-it.org (last accessed: July 22, 2011)

[7] Web Service Agreement Specification (WS-Agreement): http://www.ogf.org/documents/GFD.107.pdf (last accessed: July 22, 2011)

[8] B. Koller, Enhanced SLA Management in the High Performance Computing Domain, PhD Thesis, 2010

[9] D. Lembo, M. Lenzerini, and R. Rosati, Review on models and systems for information integration, Universita di Roma La Sapienza, 2002

[10] X. Su and J. A. Gulla, An information retrieval approach to ontology mapping. Data & Knowledge Engineering, vol. 58, pp. 47-69, 2006

[11] T. K. Landauer, P. W. Foltz, and D. Laham, An Introduction to Latent Semantic Analysis, Discourse Processes, vol. 25, pp. 259-284, 1998

[12] J. Karlgren and M. Sahlgren, "From Words to Understanding", in Y. Uesaka, P. Kanerva, and H. Asoh, Foundations of Real-World Intelligence, pp. 294-308, 2001

[13] N. Chatterjee and S. Mohan, Discovering Word Senses from Text Using Random Indexing, Proceedings CICLing, 2008

[14] M. Sahlgren, An Introduction to Random Indexing. Proceedings 7th TKE, 2005

[15] The GRAAP Working Group, web site: http://forge.gridforum.org/projects/graap-wg (last accessed: July 22, 2011)

[16] J. Zhang, Ontology and the Semantic Web, Proceedings North American Symposium on Knowledge Organization, vol. 1, 2007

[17] A. Gal and P. Shvaiko, Advances in Web Semantics I, Lecture Notes in Computer Science, vol. 4891/2009, pp. 176-198, 2009

[18] J. Huang, D. Dou, L. He, J. Dang, and P. Hayes, Ontology-based knowledge discovery and sharing in bioinformatics and medical informatics: a brief survey, Proceedings 7th Conference on Fuzzy Systems and Knowledge Discovery, pp. 2203 – 2208, 2010

[19] BonFIRE project, web site: http://www.bonfire-project.eu/ (last accessed: July 22, 2011)

 ISBN: 978-1-61208-162-5

Pervasive Ad hoc Location Sharing To Enhance Dynamic Group Tours

Markus Duchon, Corina Schindhelm
Siemens AG, Corporate Technology
CT T DE IT 1
Otto-Hahn-Ring 6, 80200 Munich
{markus.duchon.ext, corina.schindhelm}@siemens.com

Julian Köpke, Michael Dürr, Florian Gschwandtner
Ludwig-Maximilian-University Munich
Mobile and Distributed Systems
Oettingenstrasse 67, 80538 Munich
julian@die-informatiker.eu
{michael.duerr, florian.gschwandtner}@ifi.lmu.de

***Abstract*—Transportation and mobility are important factors for economics as well as social life. In this context, dynamic carsharing systems gain increasing importance, as they address traffic related problems, like increasing the occupancy rate, and therefore have a positive effect on reducing congestion. In this work, we propose a decentralized architecture and our approach to enhance these services as well as their accessibility with ad hoc location sharing. The system works based on ad hoc communication, various positioning technologies and the users' preferences. Finally, a discussion of the initial results of our prototypical implementation is presented.**

***Keywords*-Distributed information systems; Mobile ad hoc networks; Context-aware services; Transportation systems**

I. Introduction

During the past decade, the importance of Location-based Services (LBS) has increased immensely. The vast number of scenarios creates diverse services, such as simple location dependent advertisements, a finder of restaurants or other points of interests (PoI), multi-party buddy or child tracking systems, etc. Most of these systems apply a central or client server architecture like e.g. the TraX framework [1], where the client mainly determines its own location (GPS, WLAN RSSI, Cell-ID, etc.) and other service related information and sends it to a central server. The server is responsible for the computation process, the comparison of the user's location and his current interest as well as the distribution of the results. Additionally, mobility is an important factor for social life and business management and is a key element for prosperity, economic growth and the quality of life. As a result, transportation related applications enjoy great popularity. These services range from simple traveler information systems [2] to services for finding available parking spots [3] to dynamic carsharing systems [4].

In our work, we are particularly interested in addressing dynamic carsharing (also known as ad hoc ride sharing) as well as the sharing of group tickets for public and rail transport. Thus, the occupancy rate can be increased while the travel costs decreased. However, in the case of ticketsharing, it can become difficult and time consuming to search for a spare seat or offer one. Furthermore, it is quite challenging to actually find the location of a person who is willing to share a ride or ticket especially at crowded locations.

Imagine the following scenario: A group of three persons share a group ticket for up to five persons, leaving two seats unused. At the train station they decide to offer these seats. They use one of their smartphones to establish an ad hoc network and offer the remaining seats. Another person arrives at the station and before buying a single ticket he starts the TicketShare application and enters his destination. When both devices are in communication range of each other, the application searches for group offering that match the single person destination. If a match is discovered, the current location of the devices with the matches is displayed on each others device.

Therefore, with the current position transferred, it makes it possible to locate each other at the train station even though they have not met before.

By knowing the exact location of the person on the roadside, the discovery process of fellow passengers also enhances carsharing with the knowledge of whom to pick up.

A disaster management operation represent another field where ad hoc location sharing could be helpful by allowing requests for special tools or task force units to be broadcasted without the need for an existing infrastructure.

The remainder is structured as follows. Section II gives an overview of the related work. The main requirements and the resulting system architecture, including the communication and matching processes, are outlined in Section III. The prototypical implementation and preliminary results follow in Section IV. Section V concludes the paper.

II. Related Work

During the past years a lot of work has been published related to the provision of location information especially in the area of transportation and carsharing. Due to their vast number only a small overview on the most recent projects can be provided in the following.

Fu et al. [5] propose a conceptual framework for the dynamic ride sharing community on traffic grid. Thereby, users announce their travel demands as well as planed trips including the current location to a central server. The server allows for plenty additional functionalities and the route matching

 ISBN: 978-1-61208-162-5

also takes the current traffic flow and predictions into account to provide more reliable timing information for rendezvous points. However, the system utilizes a centralized approach which does not support the idea of our locally limited and distributed approach.

OpenRide [6] is a system developed by the Fraunhofer Institute which allows for spontaneous carsharing even if the users are already on the move. The transport request or carsharing offers and requests are sent to a central server which tries to find an appropriate combination of routes and passengers. In case fellow passengers are found the service responds in real-time. Although, the system supports dynamic shared rides even when the users are in good distance a central server is utilized which also does not correspond to our idea.

Piorkowski [7] sketches the idea of an application called SmartRide which utilizes short-range communication technologies and aims mostly at urban, opportunistic trips. Thereby, the author outlines the potential benefits achieved by carsharing in general. Whereas, the main challenges and requirements are discussed neither an implementation nor a specification can be found on either communication or the transfered information yet.

Winter et al. propose a peer-to-peer based shared ride trip planning system in [8]. The system operates in a distributed manner and locally exchanges relevant information. In [9] they evaluate different communication strategies and the results received by a simplified simulation indicate that the single-hop approach dramatically decreases the message overhead. But, by using a mid- or long range strategy better results regarding the overall travel time can be achieved. As the focus of their work is on the communication process including booking and revocation of trips it only addresses a part of a complete system.

Rudnicki et al. [10] propose a concept for local shared ride trip planning. They utilize a distributed approach whereby requesting clients periodically send a query within their communication range. Hosts which are coupled to vehicles answer incoming queries in case the requested route fits with the own one. The focus of this work is on the rendezvous problem in case several transfers need to be considered. However, the applied communication range of at least 1000 meters seems to be unrealistic, it could be shown by simulation that no better results can be achieved when it is further increased.

Banerjee et. at [11] present their approach for spatially restricted location exchange and implemented a Friendmeter application which calculates the users relative position by the received signal strength indication using several radio based technologies. The locations, distances and names are also displayed in the 2D euclidean space. Besides, the exchange of location data no service information or any local matching process is considered so far.

In summary, carsharing is said to reduce the load on the overall transportation system and provides several benefits. Besides some central approaches also decentralized gain more and more importance, but none of the presented approaches is able to cover the whole spectrum yet. In regard to privacy and according to Ghelawat et al. [12] people are mostly concerned about how their personal information is used by such a system and how it might be disclosed to others especially when a central server is involved.

III. System Overview

To improve carsharing or group tours by using local information exchange, we must consider some fundamental requirements. Subsequently, we present our architecture including the most important components of a distributed location sharing application in the area of group tours. This architecture can also be adapted to other scenarios, where certain service information as well as position data needs to be exchanged.

A. Requirements

With regard to privacy concerns, the responsibility for personal profiles, route matching, and the coordination of information exchanges should not be carried out by a central component. And the users themselves should have full control to what gets disclosed. Also, the dissemination within a spatially restricted area could decrease reservations, because the provided information can only be used at the current location and for a limited duration. In order to guarantee these aspects a decentralized service provisioning and application is essential. However, it should be noted that the secure transmission of information is not in the focus of this work, but will be addressed in the near future.

In order to provide an appropriate communication infrastructure with as less resources as possible the system should apply wireless communication technologies to satisfy the requirement of ad hoc communication. Therefore, generic mobile phones with integrated WLAN hardware enabling the establishment of a mobile ad hoc network seem to be sufficient.

To determine the position of each device, different technologies can be applied. Considering our use case, indoor positioning techniques should be regarded as e.g. subway stations are also targeted. Given the insufficient precision and accuracy, the solitary use of cellular positioning systems is no option. In outdoor environments with a direct line of sight, the position can be easily obtained by satellite systems like GPS. In general, the system should also allow for a decentralized location determination.

Furthermore, a decentralized and locally executable matching process is necessary to compare the transfered information in order to discover overlapping routes and profiles respectively. To match the combinability of routes, according information about the stops and their order is required too. Since our system should support several and concurrently different fellow passenger groups a scalable approach is vital. But due to the spontaneous and locally limited character we believe that a number of 200 simultaneous users is deemed to be adequate.

B. System Architecture

In favor for an autonomous ad hoc interaction between mobile devices we waive a central component and propose a

 ISBN: 978-1-61208-162-5

fully decentralized system architecture. Therefore, we decided to use modern smartphones which allow among others for ad hoc WLAN communication and offer certain positioning capabilities via GPS and other radio based hardware.

The proposed architecture is illustrated in Figure 1. The user interface handles the users input in terms of the destination and whether seats are offered or demanded. It also illustrates the applications output, e.g. if a match occurred. The application logic is responsible for the execution of some minor functions and the delegation of incoming tasks to the according subcomponents, namely `Positioning`, `Ad hoc Communication` and `Matching`. The latter two will be discussed in more detail. The positioning component detects

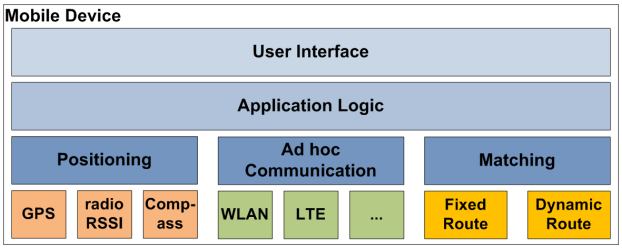
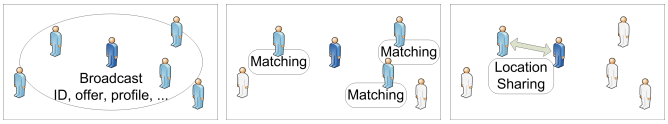


Fig. 1. System architecture for Ad hoc Location Sharing

the users exact location which can be accomplished by an integrated GPS receiver or the relative location based on the received signal strength indication using Bluetooth or WLAN, as proposed by Banerjee et al. in [11]. Also, additional sensors like a compass can be utilized as well, which in total allow for a decentralized position detection and distance measurements. The communication component has several tasks to fulfill. First, the network device needs to be configured to establish an ad hoc network functionality. Second, packets broadcasted by other clients need to be received and optionally forwarded. And third, packets which are generated by the application logic must be broadcasted. For that reason, several technologies can be applied, e.g. WLAN or LTE. The matching component compares user profiles and possible overlaps of routes. The routes can be either static like in a public transportation system or dynamic with respect to carsharing.

A simplified overview of the negotiation process is illustrated in Figure 2. In the ticketsharing scenario, a user announces a certain number of available seats, the position and orientation, the destination as well as some personal preferences within the communication range of the mobile device (Fig. 2a). Other devices and users, which receive this information locally, check if the offer fits their own needs (Fig. 2b). If so, the location and orientation of the supplier and purchaser are periodically exchanged in order to actually find each other (Fig. 2c). Especially the process of finding each other is important, so that the journey can be continued together as a group. In case that more than one seat is offered and one match has been found already, the announcement will be broadcasted again but with a decreased number of available seats. In this way we utilize the first come first serve principle.



(a) Announcement (b) Matching (c) Location Sharing

Fig. 2. Ad hoc Location Sharing Process

C. Ad hoc Communication and Protocol

To enable this kind of communication and actual data transfer using either TCP or UDP an IP address has to be assigned first. This address has to be configured by the involved devices themselves and also possible conflicts need to be resolved. A suitable protocol called APIPA [13] is utilized for this task which works as follows:

1) The client generates a random IP address within the 169.254.0.0/16 address space.
2) An ARP request for the generated address is broadcasted throughout the network.
3) If an answer is received, and the address is already in use, the process is restarted at the first step.
4) If there is no answer after a previously defined timeout of for example one second, the network adapter is configured to use the address.

For a small number of clients this protocol works good, but the full address range of $254^2 = 64516$ cannot be used efficiently, because the free addresses always have to be found by random. In our case, the chance to get a free address at the first try is still above 99% with 200 addresses in use. The communication

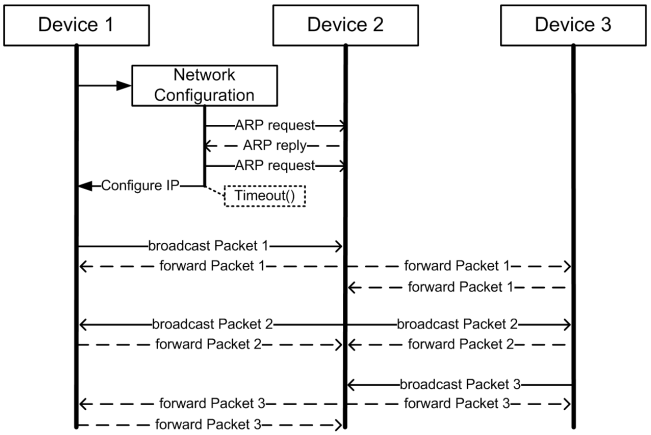


Fig. 3. Sequence diagram: network configuration and an example transmission of three packets (multihop)

between more than two devices including a multihop approach is illustrated in Figure 3. Device 2 is in the range of Device 1 and Device 3, but Device 1 is not in range of Device 3. In this example Devices 2 and 3 are already configured and the diagram starts with the configuration of Device 1 as explained earlier. By doing so, a decentralized service provisioning and application can be realized using the standard internet protocol. The service itself can be implemented by an additional protocol describing the packets to be transfered.

 ISBN: 978-1-61208-162-5

By utilizing the XML standard, as illustrated in Listing 1, the protocol is human readable, extensible, and new features can easily be added in future versions. To identify a packet,

```
<?xml version="1.0" encoding="utf-8"?>
<packet>
  <profile id="00:12:34:56:78:9A">
    <name session="1300738151790" counter="1">
          Julian
    </name>
    <settings>[...]</settings>
    <preferences>[...]</preferences>
  </profile>
  <confirmed>
    00:1F:3B:27:56:AB,00:AB:13:71:34:A3
  </confirmed>
  <origin>
    <station>Odeonsplatz</station>
    <position>
      <gps lat="48.150122" lon="11.581163"/>
      <heading>73</heading>
    </position>
  </origin>
  <destination line="U6">
    <station>Garching-Forschungszentrum</station>
    <adress/>
  </destination>
  <seats>
    <status>available</status>
    <count>3</count>
  </seats>
  [...]
</packet>
```

Listing 1. XML packet example

in addition to the device identifier, each packet will possess a basic counting ID which will start at zero and will be incremented with each new packet that is sent by that sender. Furthermore, a session ID is included, which consists of the time when the program was started. Each new packet will invalidate all previous packets by that device. This way, every device only has to store a set of device IDs together with their latest session ID and highest counting ID of that session to know if a packet has already been received or still has to be forwarded.

Within the confirmation section the device IDs are stored, which are matching with the offer. To prevent double reservations it can only be set by the supplier. This creates a certain matching overhead, but guarantees a proper distributed negotiation process. The first and vital payload of the protocol will be the position of the user in the origin section, which either consists of a pair of latitude and longitude when GPS is used, or the received signal strength indication (RSSI) of all devices in communication range in case of WLAN (`<wlan rssi1="00:1F:3B:27:56:AB,-29dBm" rssi2="..."/>`). In addition, the compass heading is included. The point of origin and the destination in form of a railway station has to be set and will be used to match the routes. Optionally, a railway line and a departure time could be selected, so different trains serving the same routes can be distinguished. Furthermore, there has to be the number of available or required seats which is vital for the proposed scenario. Other useful fields not listed are e.g. price and comment, whereby the latter one can be used to add something to the offer or request not covered yet.

The protocol also features a profile section, in which at least the user name is mandatory because it will be used to show up on other devices. The profile can be extended by certain settings and preferences like the gender, age, or habits which are used during the matching process to find suitable fellow passengers not only on the basis of the route.

D. Matching Process

First it must be guaranteed, that every user shares the same information about the network which is either a public transportation system or a general street network according to the applied use case. That information can be directly stored and handled by the mobile device as demonstrated in previous work [14]. The required information of the respective network can be easily extracted from the OpenStreetMap project [15] and preprocessed using Osmosis [16]. The map tiles provided by this project can be also utilized for the visualization and can either downloaded on demand or cached on the device beforehand.

In order to prevent double reservations the matching process first checks if the device ID is in the list of confirmed candidates and in a positive case only the name and location information is further processed. The confirmation can only be set by the supplying user to guarantee a reliable negotiation process. If not in this list, the algorithm compares the status for a supplier or a purchaser, and if the number of seats is sufficient. In case a supplier and a purchaser exchange information the process continues. Otherwise, the packet is ignored and depending on the dissemination strategy optionally forwarded. Afterwards, the profile information or preferences of the received packet are compared to the own settings. The process continues if, e.g. the age of the settings is within the interval of the desired age in the preferences, the process continues. As an alternative to the automatic comparison it would be also possible to let the users decide themselves in order to further reduce privacy concerns. By using an appropriate namespace for the settings and preferences respectively the same vocabulary is used which can be simply checked against each other. Matching and non-matching device IDs are remembered by the application and therefore are not considered by the matching process when the next packet is received.

To compare two routes, it has to be determined if both are using the same train. Thereby, and if no timetables are available, it is assumed that both take the next available train traveling to the destination. With the knowledge of the station lists which has been stored on the device the origin, destination and the railway line number can be selected by the user. The line number can also be omitted, if all stations of all lines on the route are pairwise disjoint. During the matching process it just needs to be determined if one or another destination is along the way of the other one. When a match finally occurs and is accepted by the user, the number of seats is decreased by the requested ones and the broadcast is continued. In case no seat is left the number will be zero and therefore no more match will occur, but the position data is still necessary in order to find each other.

The length of the shared route is another factor which can be accounted for. A list of matching candidates could be generated and these can be sorted by the length of the common route, whereby the seats are filled up starting with the longest one. This way, the application could not only show which

 ISBN: 978-1-61208-162-5

clients could be picked up at all, but also can make a good suggestion for whole groups to share a ride.

In the case of carsharing the exact match of a route would lead to very few results. So, instead of the route, the detour that had to be made to deliver the person to its destination is considered. Thereby, the candidates' destination is integrated as an intermediate stop and the length of the resulting route is compared with the length of the own one. If the detour is below a certain threshold, which can be configured within the settings a match occurs. This requires several route calculations and therefore a relatively fast algorithm is required, e.g. the mobile contraction hierarchy approach proposed by Sanders et al. [17]. It calculates the length of a route in the European road network in less than 60 ms and a complete routing graph in less than 100 ms.

Depending on the use case (ticket- or carsharing) one or another matching strategy is selected, whereby both allow for a local and decentralized matching process in order to support the idea of a system which works without any central component.

IV. Implementation

We implemented the presented approach using the Android platform. To enable a spontaneous interaction between devices, we utilize the WLAN ad hoc capabilities with the location being determined by the integrated GPS receiver. For testing purposes, we decided to use the ticketsharing scenario and obtain the required train network, including the order of the stations, from the OpenStreetMap project.

A. Ad hoc Communication

By the time of implementation, it was not possible with the provided Android standard tools to initiate an ad hoc WLAN connection nor possible to utilize an arping command required for our implementation of the APIPA protocol illustrated in Listing 2. Therefore, we applied some opensource modifications to a common device in order to enable the mentioned functionalities. A broadcast with an update of the location and other data is sent every second, and as a result, the GPS location is detected once per second and shared with the matching participants. First, a random IP address is generated with the

```
   private String generateIP() {
2     Random rand = new Random();
      return ip = "169.254." + (rand.nextInt(254) + 1)
4               + "." + (rand.nextInt(254) + 1);
   }
6
   private void checkIP() {
8     try {
         su.writeBytes("busybox arping "
10       + "-D -f -c 1 -w 1 -I `getprop wifi.interface` "
         + ip + "\n");
12    } catch (IOException e) { e.printStackTrace(); }
   }
14
   private void setIP() {
16    try {
         su.writeBytes("export brncl_lan_gw=" + ip + "\n");
18       su.writeBytes(TicketShare.getInstance()
         .getFilesDir() + "/wifi config\n");
20    } catch (IOException e) { e.printStackTrace(); }
   }
```

Listing 2. Implementation of the APIPA example

`generateIP` method, then it is checked using the `checkIP` method and if no response is received after a timeout of 1 second, the interface is configured by the `setIP` method. To create an actual network connection without any centralized configuration, there have to be some fixed variables shared by all participating devices: (a) The SSID of the wireless network was predefined to `TicketShare`, but regarding the use case also other names can be used. (b) The wireless channel defines the frequency range on which the radio device sends and receives packets and is set to 1. (c) The UDP port is a virtual identifier for the service implemented by the protocol. Since port 31337 is not used by any known application, it was selected for the whole communication process.

To increase the coverage, we also implemented the proposed multihop broadcasting, whereby the number of hops could be limited by the time to life (TTL). Every unique packet received is sent out again by the device until the TTL value is zero. To identify a packet and to decide which one needs to be retransmitted the device ID, the session ID, and the counting ID were used. This can dramatically increase the coverage, especially when there is a high density of devices like e.g. at train stations.

B. Matching

For testing purpose, we extracted the subway network of Munich, Germany from the OpenStreetMap project and stored the resulting, ordered stations as XML files (see Listing 3) on every device, because for an according matching process all clients have to share the same network information. Otherwise, there could be inconsistencies between the matching results shown on different devices that must be avoided.

```
1  <?xml version="1.0" encoding="utf-8"?>
   <routes>
3     <stations route="U1">
         <station name="Olympia-Einkaufszentrum" />
5        <station name="Georg-Brauchle-Ring" />
         <station name="Westfriedhof" />
7        <station name="Gern" />
         <station name="Rotkreuzplatz" />
9        [...]
         <station name="Mangfallplatz" />
11    </stations>
      <stations route="U2">
13       [...]
      </stations>
15    [...]
   </routes>
17 }
```

Listing 3. Example of an XML station list

First, the device ID is looked up in a list of already matching candidates. If not in the list, the algorithm checks the availability of the ticket, otherwise the name and location information is used to display the candidate on a map.

The cases where both users provide a ticket or neither offers a ticket represent a non-match. When only one of them has a ticket, the number of available seats is compared with the required number. If more or equal seats are available than needed, the algorithm continues, otherwise, they do not match.

Then the preferences of the profile are matched with the own settings and in the positive case the algorithm continues, otherwise no match was found and the device ID is stored in another list of non-matching candidates. In the implemented use case the next field being matched is the line. If this is not equal, the users will travel in different trains and therefore will not be able to share a ticket. Finally, the destinations

 ISBN: 978-1-61208-162-5

are matched and if they are on the same route the actual destination of the supplier is irrelevant in this context, because it is assumed that the person with the longest route takes the ticket. If the route can be traveled together a match is found and the according device ID is remembered and the according information is used to display the name and location of the candidate. Otherwise no match could be found at all but the device ID is also remembered in order the matching process will not be repeated when following packets are received.

C. Discussion

We tested our implementation with four devices communicating simultaneously and equal origins heading to different destinations. The decentralized service provisioning and application worked and matching candidates were displayed including their current locations and distances of every involved device. Due to the interval of one second per broadcast the location updates were sufficient and the participants were able to find each other easily. Only the orientation information was subject to high deviations depending on the integrated hardware. The matching process for the implemented use case was also below the time of a location update and without appreciable delays. The maximum distance that could be achieved using a single hop mechanism covered about 150 meters with a direct line of sight and could be extended by the multihop broadcasting approach.

With our limited number of devices we were not able to test the maximum capacity of ad hoc networks, but by calculation and according to the analysis of Li et al. in [18] the network will be clogged with 130 devices broadcasting simultaneously when using an unlimited forwarding strategy and a packet size of around 400 bytes, which is about the size of the example packet in Listing 1.

Because of the very fast route length calculation from the work of Sanders et al. [17] also the carsharing use case seems feasible. When riding towards a group of potential fellow passengers at a speed of 30 km/h and assuming a reduced communication range of 75 meters about 150 route length calculations are possible in theory until the car passes the group which is largely enough.

In summary our approach of a decentralized ad hoc location sharing could definitely enhance group tours and their accessibility. Certainly, ulterior and comprehensive tests are necessary and will be conducted in future work.

V. Conclusion

In this paper we presented a decentralized approach to enhance alternative modes of transportation, such as ticket- and carsharing, as well as their accessibility. By utilizing ad hoc communication, a distributed matching approach, and terminal based positioning technologies, we were able to demonstrate the feasibility of our service based on commonly used smartphones. Additional improvements regarding social aspects like the ranking of fellow passengers and price negotiations are also planned. Furthermore, different position approaches need to be evaluated in more detail. Given that the focus of this work was the proof of concept for a decentralized location sharing approach, future work will concentrate on a secure authentication and transmission of necessary service information. The use of IPv6 as communication protocol instead of IPv4 could leverage the network configuration process because each device can be addressed directly. Finally, the ad hoc capabilities of LTE network and the possibility to integrate this technology into our prototype remain interesting topics.

References

[1] A. Küpper, G. Treu, and C. Linnhoff-Popien, "TraX: a device-centric middleware framework for location-based services," *IEEE Communications Magazine*, vol. 44, no. 9, pp. 114–120, 2006.

[2] B. Ferris, K. Watkins, and A. Borning, "Location-Aware Tools for Improving Public Transit Usability," *IEEE Pervasive Computing*, vol. 9, pp. 13–19, 2010.

[3] C. J. Rodier and S. A. Shaheen, "Transit-based smart parking: An evaluation of the San Francisco Bay area field test," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 2, pp. 225 – 233, 2010.

[4] J. Shao and C. Greenhalgh, "DC2S: a dynamic car sharing system," in *Proceedings of the 2nd ACM SIGSPATIAL*, ser. LBSN '10. New York, NY, USA: ACM, 2010, pp. 51–59.

[5] Y. Fu, Y. Fang, C. Jiang, and J. Cheng, "Dynamic Ride Sharing Community Service on Traffic Information Grid," *Intelligent Computation Technology and Automation, International Conference on*, vol. 2, pp. 348–352, 2008.

[6] Fraunhofer FOKUS, "OpenRide," 2009, last accessed 27. July 2011. [Online]. Available: http://www.open-ride.com/english/index.php

[7] M. Piorkowski, "Collaborative Transportation Systems," in *WCNC, 2010 IEEE*, 2010, pp. 1–6.

[8] S. Winter, S. Nittel, A. Nural, and T. Cao, "Shared Ride Trip Planning with Geosensor Networks," in *Institute of Geoinformatics, University of Muenster*, 2005, pp. 135–146.

[9] S. Winter and S. Nittel, "Ad hoc shared ride trip planning by mobile geosensor networks," *International Journal of Geographical Information Science*, vol. 20, no. 8, pp. 899–916, 2006.

[10] R. Rudnicki, K.-H. Anders, and M. Sester, "Rendezvous-Problem in Local Shared-Ride Trip Planning," International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVIII, 2008.

[11] N. Banerjee, S. Agarwal, P. Bahl, R. Chandra, A. Wolman, and M. Corner, "Virtual Compass: Relative Positioning to Sense Mobile Social Interactions," in *Pervasive Computing*. Springer Berlin / Heidelberg, 2010, vol. 6030, pp. 1–21.

[12] S. Ghelawat, K. Radke, and M. Brereton, "Interaction, privacy and profiling considerations in local mobile social software: a prototype agile ride share system," in *Proceedings of the 22nd Conference of the Computer-Human Interaction*, ser. OZCHI '10. New York, NY, USA: ACM, 2010, pp. 376–379.

[13] C. M. Kozierok. (2003) APIPA - Automatic Private IP Addressing. [Online]. Available: http://www.tcpipguide.com/free/t_DHCPAutoconfigurationAutomaticPrivateIPAddressingA-2.htm

[14] M. Duchon, A. Paulus, and M. Werner. Mobile Anwendung zur Routenplanung mit öffentlichen Verkehrsmitteln basierend auf OpenStreetMap-Daten. FOSSGIS 2011.

[15] S. Coast. (2004) OpenStreetMap - The Free Wiki World Map. Last accessed 27. July 2011. [Online]. Available: http://www.openstreetmap.org

[16] Osmosis. Processing OSM data. Last accessed 27. July 2011. [Online]. Available: http://wiki.openstreetmap.org/index.php/Osmosis

[17] P. Sanders, D. Schultes, and C. Vetter, "Mobile Route Planning," in *Algorithms - ESA 2008*, ser. Lecture Notes in Computer Science, D. Halperin and K. Mehlhorn, Eds. Springer Berlin, Heidelberg, 2008, vol. 5193, pp. 732–743.

[18] J. Li, C. Blake, D. S. J. D. Couto, H. I. Lee, and R. Morris, "Capacity of Ad Hoc Wireless Networks," in *7th ACM International Conference on Mobile Computing and Networking*, 2001.

 ISBN: 978-1-61208-162-5

Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs

Jingjing Liu, Alice Li and Stephanie Seneff
MIT Computer Science & Artificial Intelligence Laboratory
Cambridge, MA 02139, USA
{jingl, aliceli, seneff@csail.mit.edu}

Abstract—**In recent years, consumers have become empowered to share personal experiences regarding prescription drugs via Web page discussion groups. This paper describes our recent research involving automatically identifying adverse reactions from patient-provided drug reviews on health-related web sites. We focus on the statin class of cholesterol-lowering drugs. We extract a complete set of side effect expressions from patient-submitted drug reviews, and construct a hierarchical ontology of side effects. We use log-likely ratio estimation to detect biases in word distributions when comparing reviews of statin drugs with age-matched reviews of a broad spectrum of other drugs. We find a highly significant correlation between statins and a wide range of disorders and conditions, including diabetes, amyotrophic lateral sclerosis (ALS), rhabdomyolysis, neuropathy, Parkinson's disease, arthritis, memory loss, and heart failure. A review of the research literature on statin side effects corroborates many of our findings.**

Keywords- medicine data mining; drug side effect discovery

I. INTRODUCTION

The last few decades have witnessed a steady increase in drug prescriptions for the treatment of biometric markers rather than overt physiological symptoms. Today, people regularly take multiple drugs in order to normalize serum levels of biomarkers such as cholesterol or glucose, or to reduce blood pressure. All drugs have side effects, which are sometimes debilitating or even life-threatening. When a person taking multiple drugs experiences a new symptom, it is not always clear which, if any, of the drugs or drug combinations are responsible.

Increasingly, consumers are turning to the Web to seek information, and, increasingly, this information comes in the form of consumer-provided comments in discussion groups or chat rooms. User reviews of products and services have empowered consumers to obtain valuable data to guide their decision process. Recently, statistical and linguistic methods have been applied to large datasets of reviews to extract summary and/or rating information in various domains ([9] [22]).

Health care and prescription drugs are a growing topic of discussion online, not surprising given that almost half of all Americans take prescription drugs each month, costing over $200 billion in 2008 alone ([5]). Though drugs are subject to clinical trials before reaching market, these trials are often too short, and may involve too few people to give conclusive results. A large study recently conducted on the heart failure drug, nesiritude, invalidated the findings of the smaller study that had led to the drug's approval [11]. While regulatory agencies do attempt to monitor the safety of approved medical treatments, surveillance programs such as the U.S. Food and Drug Administration's (FDA's) Adverse Event Reporting System (AERS) are often difficult for patients to use.

In addition, the large language gap between medical documents and patient vocabulary can cause confusion and misunderstanding ([23]). We hope to take advantage of the vast amount of information available in patient anecdotes posted online to address the dual problems of insufficient clinical studies and mismatched terminology.

We envision a system that increases patient awareness of drug-related side effects by enabling consumers of prescription drugs to easily browse a large consolidated database of posts from health-related web sites. Beyond aggregating data from drug review and health discussion sites, we plan to support spoken queries, which would be answered via a set of succinctly summarized hits that best match the query, based on sophisticated statistical and linguistic techniques. The user could then click on any one of these displayed summaries to read the associated post.

This paper describes our preliminary efforts to detect associations between a drug class and its side effects. We use statistics and heuristic methods to build up a hierarchical ontology of side effects by aggregating patient-submitted drug reviews. We use log-likelihood ratios to extract summary information derived from biases in word and phrase distributions, and to quantify associations between drugs and symptoms. For the scope of this paper, we focus on statin drugs, which are among the most costly and commonly prescribed drugs in the United States. The methods described are applicable to all drug classes.

In the remainder of this paper, we will first review the research literature reflecting known or suspected side effects associated with statin drugs. After explaining our data collection and side-effect ontology construction, we describe our methodology and verify that many of our extracted associations align with observations from the literature.

II. BRIEF LITERATURE REVIEW

A. Side Effects of Statin drugs

Statins (Hydroxy methyl glutaryl coenzyme A reductase inhibitors) have become increasingly popular as very

 ISBN: 978-1-61208-162-5

effective agents to normalize serum cholesterol levels. The most popular of these, atorvastatin, marketed under the trade name, Lipitor, has been the highest revenue branded pharmaceutical for the past 6 years. The official Lipitor web site lists as potential side effects mainly muscle pain and weakness and digestive problems. However, several practitioners and researchers have identified suspected side effects in other more alarming areas, such as heart failure, cognition and memory problems, and even severe neurological diseases such as Parkinson's disease and ALS (Lou Gehrig's disease). [21] provides compelling arguments for the diverse side effects of statins, attributing them mainly to cholesterol depletion in cell membranes.

It is widely acknowledged that statin drugs cause muscle pain, weakness and damage ([7] [12]), likely due in part to their interference with the synthesis of the potent antioxidant Coenzyme Q10 (CoQ10) ([10]). CoQ10 plays an essential role in mitochondrial function to produce energy. Congestive heart failure is a condition in which the heart can no longer pump enough blood to the rest of the body, essentially because it is too weak. Because the heart is a muscle, it is plausible that heart muscle weakness could arise from long-term statin usage. Indeed, atorvastatin has been shown to impair ventricular diastolic heart performance ([14]). Furthermore, CoQ10 supplementation has been shown to improve cardiac function ([13] [20]).

The research literature provides plausible biological explanations for a possible association between statin drugs and neuropathy ([15] [24]). A recent evidence-based article ([1]) found that statin drug users had a high incidence of neurological disorders, especially neuropathy, parasthesia and neuralgia, and appeared to be at higher risk to the debilitating neurological diseases, ALS and Parkinson's disease. The evidence was based on careful manual labeling of a set of self-reported accounts from 351 patients. A mechanism for such damage could involve interference with the ability of oligodendrocytes, specialized glial cells in the nervous system, to supply sufficient cholesterol to the myelin sheath surrounding nerve axons. Genetically-engineered mice with defective oligodendrocytes exhibit visible pathologies in the myelin sheath which manifest as muscle twitches and tremors ([16]).

Cholesterol depletion in the brain would be expected to lead to pathologies in neuron signal transport, due not only to defective myelin sheath but also to interference with signal transport across synapses ([17]). Cognitive impairment, memory loss, mental confusion, and depression were significantly present in Cable's patient population ([1]). Wagstaff et al. ([19]) conducted a survey of cognitive dysfunction from AERS data, and found evidence of both short-term memory loss and amnesia associated with statin usage. Golomb et al. ([6]) conducted a study to evaluate evidence of statin-induced cognitive, mood or behavioral changes in patients. She concluded with a plea for studies that "more clearly establish the impact of hydrophilic and lipophilic statins on cognition, aggression, and serotonin."

B. Relationship between Cholesterol and Health

ALS and heart failure are both conditions for which published literature suggests an increased risk associated with statin therapy ([1] [10]). Indeed, for both of these conditions, a survival benefit is associated with elevated cholesterol levels. A statistically significant inverse correlation was found in a study on mortality in heart failure. For 181 patients with heart disease and heart failure, half of those whose serum cholesterol was below 200 mg/dl were dead three years after diagnosis, whereas only 28% of the patients whose serum cholesterol was above 200 mg/dl had died. In another study on a group of 488 patients diagnosed with ALS, serum levels of triglycerides and fasting cholesterol were measured at the time of diagnosis ([2]). High values for both lipids were associated with improved survival, with a p-value <0.05.

A very recent study on the relationship between various measures of cholesterol status and health in the elderly came up with some surprising results, strongly suggesting that elevated cholesterol is beneficial for this segment of the population [18]. A study population initially over 75 years old was followed over a 17 year period beginning in 1990. In addition to serum cholesterol, a biometric associated with the ability to synthesize cholesterol (lathosterol) and a biometric associated with the ability to absorb cholesterol through the gut (sitosterol) were measured. For all three measures of cholesterol, low values were associated with a poorer prognosis for frailty, mental decline and early death. A reduced ability to *synthesize* cholesterol showed the *strongest* correlation with poor outcome. Individuals with high measures of all three biometrics enjoyed a 4.3 year extension in life span, compared to those for whom all measures were low.

III. SIDE-EFFECT DISCOVERY

A. Data Collection

To learn the underlying associations between side effects and drug usage from patient-provided reviews, we collected drug reviews from three drug discussion forums ("AskAPatient.com," "Medications.com" and "WebMD.com") which allow users to post reviews on specific drugs and share their experiences. Table 1 gives the statistics on the review data collection. A total of 8,515 statin reviews were collected from the three data sources. We also collected 105K drug reviews from the AskAPatient.com, on drugs to treat a broad range of problems such as depression, acid reflux disease, high blood pressure, diabetes, etc. This set includes reviews for non-statin cholesterol lowering drugs.

Table 1. Statistics on drug review data collection.

Data source	Number of Statin reviews
AskAPatient.com	2,647
Medications.com	4,162
WebMD.com	1,706
Total	8,515

A typical review entry contains the personal information of the user (e.g., gender, age), the dosage and duration of the drug treatment, the reason for taking the drug, the side effects that the user has experienced, as well as a free-style text comment. An example of a review is shown in Figure 1.

:drug "Lipitor"
:dosage "40mg 1X D"
:sex "Male"
:age "47"
:duration "4 years"
:reason "high cholesterol"
:side_effects "*Body aches, joint pain, decreased mobility, decreased testosterone and libido, difficulty getting out of bed in the morning, tingling and itchy hands, and decrease in overall strength.*"
:comment "I have been taking lipitor for many years. I started out on 10mgs and now I am on 40mgs. I have had hip replacement, back surgery, and shoulder surgery while on this drug. I have seen my strength decrease dramatically …"

Figure 1. Example of a review from AskAPatient.com.

B. Side-Effect Extraction

Most previous medical natural language processing research relies on medical lexicons such as those provided by Unified Medical Language System (UMLS) or the Food and Drug Administration's (FDA) COSTART corpus. However, these official lexicons often have low coverage of colloquial side effect expressions, which are very common in patient-submitted reviews. Thus, in this study we extract side effect expressions from the reviews themselves, instead of using these restrictive lexicons.

As shown in the example in Figure 1, the input of "side effects" often contains a list of short phrases describing the major side effects the reviewer has, and the input of "comment" contains free-style texts which tell the story about the reviewer's experience with the drug. To obtain a clean set of side effect expressions, we first automatically extract short phrases from the input of "side_effects" in each review entry. From the 107K reviews collected from AskAPatient.com (including both statin and non-statin drugs), we extracted 7,500 words and phrases that describe common side effects on various drugs.

These phrases extracted from general users' input contain a lot of noise. For example, some users may type in long sentences describing their conditions instead of using short phrases. Also, many phrases may describe the same type of side effect (e.g., "joint pain," "pain in joint" and "severe pain in the arm and leg joint"). To eliminate the noise and redundancy, the extracted phrases were first subjected to a stop-word filter, eliminating 377 common stop words. A phrase which contains only stop words is filtered out as noise. We also filter out the phrases which have frequency counts less than five in the whole review dataset. We further filter the phrases by grouping phrases containing the same set of non-stop-word (e.g., "joint pain" and "pain in joint"). With this filtering process, the number of candidate side effect phrases shrinks to 2,314.

C. Side-Effect Onology

To organize the set of side effect phrases, we asked an annotator who is knowledgeable in medical terminology to classify the phrases into a hierarchical ontology. First, synonyms are identified and grouped (as shown in Table 2). For example, "elevated blood pressure," "increase in blood pressure" and "higher blood pressure" are clustered into the same group. Then, these synonym groups are further organized into broad classes. For example, the synonym groups of "achy legs," "muscle pain" and "joint pain" are all clustered into the generic class of "pain."

Table 2. Examples of the synonyms of side effects.

Group	Synonyms
loss of mental clarity	mental slowness, slow brain, fuzzy thinking, foggy head, cloudy head, muddled thinking
all body aches	achy body, achy feeling, achy bones, achy all over, overall aches, body ache, aches and pains
forgetting words	difficulty finding the right word, mixing up words, can't find words, difficulty finding words
diabetes	diabetic, high blood sugar, elevated blood sugar

As a result of this manual process, the 2,314 side effect phrases are clustered into 307 synonym groups, which are furthered grouped into 30 classes. Table 3 shows the 30 classes as well as the number of synonym groups in each class, and Table 4 gives examples of the synonym groups in some classes. Note that this classification schema encompasses side effects for all drugs, and thus can be used for other drug classes besides statins.

Table 3. Classes of side effects.

Class	#Syn. groups	Class	#Syn. groups	Class	#Syn. groups
aches	11	eyes	6	mouth	11
appetite	6	hair	4	muscle	11
arthritis	4	heart	10	nerve	24
blood	9	infection	10	pain	29
breasts	4	kidney	12	skin	15
breathing	5	libido	9	sleep	13
cognition	13	liver	4	swelling	12
conditions	10	menstrual	8	taste	3
digestion	20	mobility	5	temperature	3
ears	4	mood	27	weight	5

Table 4. Example groupings of side effects into classes.

Class	Synonym groups
cognition	brain shocks, clearer thinking, cognitive problems, dementia, loss for words, loss of mental clarity, memory problems, mental instability, problems concentrating, short attention span
heart	atrial fibrillation, heart attack, heart failure, heart valve, heart palpitations, high heart rate, high pulse, low heart rate, tightness in chest, potassium
mood	aggressive behavior, anxiety, bipolar, bizarre thoughts, blunted emotions, crying easily, depression, despair, disoriented, euphoria
muscle	fatigue, loss of muscle mass, loss of muscle tone, muscle cramps, muscle pain, muscle spasms, muscle tightness, muscle weakness, rhabdomyolysis

 ISBN: 978-1-61208-162-5

D. Association of Drug Class with Side-Effects

Given the hierarchical ontology of side effects, we can now discover which side effects are strongly associated with statin drugs. For this, we make use of log-likelihood ratio ([3]). In statistics, a likelihood ratio test is used to compare the fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a *p*-value to decide whether to reject the null model in favor of the alternative model.

To apply the log likelihood ratio algorithm, we treat the side effect association problem as a coin toss model. The set of reviews on statin drugs (R_1) is analogous to a coin *A*. The set of reviews on non-statin drugs (R_2) is analogous to a coin *B*. Each review (in R_1 or R_2) is a coin toss instance. For a certain side effect phrase t, if a review contains the phrase, the "Head" of the coin shows up; otherwise, the "Tail" shows up. Thus, the null hypothesis is that coin *A* and coin *B* have the same probability of showing "Head" or "Tail," i.e., the review set R_1 and R_2 have the same probability of containing the phrase. The alternative model is that coin *A* and coin *B* have different probabilities of showing up "Head" or "Tail," i.e., one review set (R_1 or R_2) has a higher probability of containing the phrase than the other. The measurement of the hypothesis that the phrase t is more likely to occur in the set of statin reviews (R_1) is calculated by:

$$L_1 = k_1 log \frac{p_1}{p} + (n_2 - k_2) log \frac{1-p_2}{1-p} \quad (1)$$

where k_1 is the counts of statin reviews that contain the side effect phrase t, k_2 is the counts of non-statin reviews that contain the phrase t, p_1 is the probability of the phrase t occuring in R_1, p_2 is the probability of the phrase t occuring in R_2, p is the probability of the phrase t occurring in the whole document set ($R_1 \cup R_2$), n_1 is the size of R_1, and n_2 is the size of R_2.

Maximum likelihood is achieved by:

$$p_1 = \frac{k_1}{n_1} \qquad p_2 = \frac{k_2}{n_2} \quad (2)$$

$$p = \frac{k_1+k_2}{n_1+n_2} \quad (3)$$

A symmetric equation of (1) can be derived for L_2, the hypothesis that the phrase t occurs more frequently in R_2. Whether the alternative model fits significantly better and should thus be preferred can be determined by deriving the probability or *p*-value of the obtained difference $L_1 - L_2$. The probability distribution of the difference can be approximated by a chi-square distribution. *P*-values are computed under the assumption that there is one degree of freedom between the null model and the alternative model.

IV. EXPERIMENTS

In our dataset of reviews, the size of non-statin reviews (105K) is much larger than that of statin reviews (8,515). To make the two document sets equivalent for comparison, we randomly select the same size of reviews (8,515) from the non-statin reviews as R_2. An important consideration is to correct for a possible age bias of review-providers in the data selection process. Figure 2 gives the age distribution of statin drug reviewers. We observed that, in the statin review set, most of the reviews (83.6%) are published by users aged from 40 to 70. To avoid the possible bias on symptoms of different ages, we follow the same distribution of reviewers' age for the random selection of non-statin reviews.

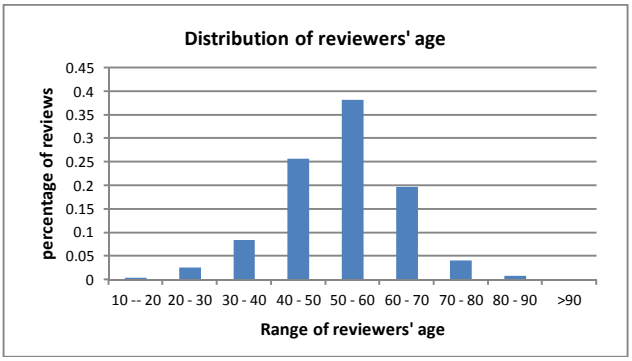


Figure 2. Distribution of reviewers' age in statin reviews.

For each side effect in our hierarchy ontology, we calculate the log likelihood ratio of the statin review set and non-statin review set as explained in the algorithm section. We treat all the synonyms for each side-effect equally, i.e., the occurrences of alternatives in the same group count for the same phrase t.

Table 5 lists all side effect clusters related to pain that yielded a *p*-value less than 0.05. Pain in essentially all parts of the body -- arms, legs, neck, shoulder, and back, all occurred more frequently in the statin reviews by a substantial margin. "Muscle pain" in particular is overwhelmingly associated with statins, with a *p*-value of 1.4E-06.

Table 5: Pain in various parts of the body (sorted by *p*-value).

Side effect	k_1	k_2	$L_1 - L_2$	*p*-value
muscle pain	1029	221	1419.26	1.4E-06
pain	2499	1557	1444.31	0.00004
pain in legs	570	265	514.16	0.00114
shoulder pain	142	34	189.65	0.00485
back pain	323	163	265.20	0.00738
neck pain	111	36	130.77	0.01417
pain in arms	76	21	96.38	0.02009

Table 6 provides the counts and *p*-values for a number of side effect clusters associated with muscle frailty and pathology. Highly disturbing is the very low *p*-value for "difficulty walking" (0.0004). Rhabdomyolysis is a frequently fatal condition involving kidney failure due to toxic exposure to myoglobin debris released into the blood stream following muscle breakdown. Since it is generally

 ISBN: 978-1-61208-162-5

rare, it did not occur at all in the non-statin reviews, but appeared 31 times in the statin reviews. "Loss of muscle mass" is ten times as frequent in the statin reviews. "Muscle cramps," "general weakness" and "muscle weakness," highly associated with "frailty," have extremely low *p*-values. In addition, "general numbness" and "muscle spasms" are also significantly associated with statins.

Table 6: Muscle frailty and pathology pain (sorted by p-value).

Side effect	k_1	k_2	$L_1 - L_2$	*p*-value
muscle cramps	678	193	850.12	0.00005
general weakness	687	210	834.24	0.00006
muscle weakness	302	45	448.73	0.00023
difficulty walking	419	128	508.96	0.00044
loss of muscle mass	54	5	84.75	0.01332
general numbness	293	166	203.34	0.01552
muscle spasms	136	57	135.03	0.01849
rhabdomyolysis	31	0	51.52	0.02177
tendonitis	42	8	59.68	0.03193
balance problems	71	32	65.91	0.05371

Table 7 shows the frequency distributions for several often debilitating conditions associated with pathologies in the brain and nervous system. Most alarming to us is the 10:1 ratio of incidence of ALS, a debilitating disease associated with damaged motor neurons in the spinal cord that is nearly always fatal. The associated *p*-value of 0.008 makes this result highly significant. The ratio is even higher for Parkinson's disease (18:1). Parkinson's disease involves damage to dopamine-secreting cells in the substantia nigra. The *p*-value for memory problems is also very low (0.01), providing powerful evidence that statins cause memory problems. An extreme form of memory problems, dementia, comes in with a *p*-value just above the significance cutoff at 0.056. Neuropathy, due to nerve damage in the peripheral nervous system, is generally associated with muscle weakness, cramps, and spasms, other side effects that occur very frequently in statin drug reviews.

Table 7: Issues related to brain and nervous system.

Side effect	k_1	k_2	$L_1 - L_2$	*p*-value
ALS	71	7	110.75	0.00819
memory problems	545	353	286.76	0.01118
Parkinson's disease	53	3	85.38	0.01135
neuropathy	133	73	97.03	0.04333
dementia	41	13	48.80	0.05598

Table 8 shows other major health issues for which the word frequencies are highly skewed towards the statin reviews. Most of these distributions are highly significant, with a p-value < 0.01. Diabetes is especially striking, with three times the frequency of occurrence in the statin reviews as in the other reviews, despite the fact that diabetes medications are included in the other class. The highly significant results for diabetes are in line with recent concern about the possibility that statins may increase risk to diabetes ([4] [8]).

"Heart attack" has an extremely low *p*-value, but in this case strong compounding from a precondition is undeniable. A similar issue arises with "stroke." The strongest correlation among the remaining conditions is found for liver damage (p<0.003), a potential side effect that is acknowledged by the statin manufacturers. It is interesting that arthritis associates strongly with statins, as arthritis has not been identified as a known side effect. Also, "heart failure" and "raised liver enzymes" are under the cutoff of 0.05, and "kidney failure" is six times as frequent in the statin reviews, with a *p*-value slightly above 0.05.

Table 8: Various other conditions.

Side effect	k_1	k_2	$L_1 - L_2$	*p*-value
heart attack	299	73	396.87	0.00068
liver damage	326	133	331.15	0.00285
diabetes	185	62	214.2	0.00565
stroke	147	44	180.18	0.00700
arthritis	245	120	208.51	0.01117
raised liver enzymes	61	22	67.52	0.04204
heart failure	36	8	49.17	0.04473
kidney failure	26	4	38.56	0.05145

To learn the high level association between side effects and statin drugs, we further aggregate the side effects into classes, and calculate the log likelihood ratio as well as the *p*-values for each class. Table 9 gives the top-ranked classes with *p*-values below 0.05 for statin reviews. These categories are considered as most strongly associated with statin drugs. In particular, "muscle problems" is overwhelmingly associated with statins, with a *p*-value of 2.0E-07.

Table 9. Side effect classes associated with statin drugs.

Class of side effect	k_1	k_2	$L_1 - L_2$	*p*-value
muscle problems	4188	2060	3549.73	2.0E-07
mobility	535	199	581.47	0.00049
liver problems	404	163	413.99	0.00166
pain	4735	3908	731.07	0.00308
nerve problems	1196	894	380.06	0.01108
arthritis	456	317	194.72	0.02690

V. DISCUSSION

In this paper, we have described our vision of a Web-based database providing potential users with a rich facility for exploring the association of prescription drugs with possible side effects. We used the basic strategy of comparing word frequency distributions between two databases as a means to uncover statistically salient phrase patterns. Our efforts focused on statin drugs, as these are a widely prescribed medication with diverse side effects. Through standard statistical log likelihood ratio estimation, we have shown that statin drugs are very strongly associated with muscle pain and weakness, and that there is as well a statistically significant association between statin drugs and several debilitating diseases, such as ALS, Parkinson's disease, rhabdomyolysis, and heart failure. Many of our findings are supported by the research literature on statins.

Our research was inspired by the study conducted by Jeff Cable ([1]). While he looked at only 350 reviews, he used careful manual analysis to deduce associated side effects. We

 ISBN: 978-1-61208-162-5

looked at a much larger set of reviews (over 8,000), and used statistical techniques for analysis. On the one hand, it is gratifying that both methods uncovered similar side-effect profiles on different data. On the other hand, it is disturbing that a drug class as widely prescribed as the statin drugs has such severe and sometimes life-threatening adverse reactions.

One limitation of the method is the compounding effects of preconditions. This clearly influences the bias for statins with regard to "heart attack" and "stoke," but may also contribute to other terms such as diabetes and heart failure. In addition, users occasionally post comments that discuss potential side effects that they did not personally experience.

VI. FUTURE WORK

In the future, we will focus on incorporating the results of our statistical analyses into the user database. We will also develop techniques to summarize individual reviews and provide associated index terms. An ambitious goal is to use parsing techniques to extract a story line that captures cause-and-effect relationships. For instance, by commenting, "It's only been 2 days without the medication and cramping is improving," a user clearly implied that the drug had caused the cramping. We also plan to expand the database to other drug classes, such as psychopharmaceuticals and acid reflux therapies. Finally, we would like to provide a speech-based interface for querying the database.

ACKNOWLEDGMENT

This research is supported by Quanta Computers, Inc. through the T-Party project.

REFERENCES

[1] J. Cable. 2009. Adverse Events of Statins - An Informal Internet-based Study. JOIMR, 7(1).

[2] J. Dorstand, P. Kühnlein, C. Hendrich, J. Kassubek, A.D. Sperfeld, and A.C. Ludolph. 2010. Patients with elevated triglyceride and cholesterol serum levels have a prolonged survival in amyotrophic lateral sclerosis. J Neurol, in Press: Published online Dec. 3 2010.

[3] T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics, 19(1):61–74.

[4] M.R. Goldstein and L. Mascitelli. 2010. Statin-induced diabetes: perhaps, it's the tip of the iceberg. QJM, Published online, Nov 30.

[5] Q. Gu, C.F. Dillon, and V.L. Burt. 2010. Prescription drug use continues to increase: U.S. Prescription drug data for 2007-2008. NCHS Data Brief, (42):1.

[6] B.A. Golomb, M.H. Criqui, H. White, and J.E. Dimsdale. 2004. Conceptual foundations of the UCSD Statin Study: a randomized controlled trial assessing the impact of statins on cognition, behavior, and biochemistry. Archives of Internal Medicine, 164(2):153.

[7] J. Hanai, P. Cao, P. Tanksale, S. Imamura, E. Koshimizu, J. Zhao, S. Kishi, M. Yamashita, P.S. Phillips, V.P. Sukhatme, et al. 2007. The muscle-specific ubiquitin ligase atrogin-1/MAFbx mediates statin-induced muscle toxicity. Journal of Clinical Investigation, 117(12):3940–3951.

[8] J. Hagedorn and R. Arora. 2010. Association of Statins and Diabetes Mellitus. American Journal of Therapeutics, 17(2):e52.

[9] J. Liu and S. Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In Proc. EMNLP, 161–169. Association for Computational Linguistics.

[10] P.H. Langsjoen and A.M. Langsjoen. 2003. The clinical use of HMG CoA-reductase inhibitors and the associated depletion of coenzyme Q10. A review of animal and human publications. Biofactors, 18(1):101–111.

[11] R. Califf. 2010. LBCT I, Abstract 21828. Presented at: American Heart Association Scientific Sessions, Nov. 13-17; Chicago.

[12] M.G. Mohaupt, R.H. Karas, E.B. Babiychuk, V. Sanchez-Freire, K. Monastyrskaya, L. Iyer, H. Hoppeler, F. Breil, and A. Draeger. 2009. Association between statin-associated myopathy and skeletal muscle damage. Canadian Medical Association Journal, 181(1-2):E11.

[13] S.L. Molyneux, C.M. Florkowski, A.M. Richards, M. Lever, J.M. Young, and P.M. George. 2009. Coenzyme Q10; an adjunctive therapy for congestive heart failure? Journal of the New Zealand Medical Association, 122:1305.

[14] M.A. Silver, P.H. Langsjoen, S. Szabo, H. Patil, and A. Zelinger. 2004. Effect of atorvastatin on left ventricular diastolic function and ability of coenzyme Q10 to reverse that dysfunction. The American Journal of Cardiology, 94(10):1306–1310.

[15] C. Silverberg. 2003. Atorvastatin-induced polyneuropathy. Annals of Internal Medicine, 139(9):792.

[16] G. Saher, B. Brügger, C. Lappe-Siefke, W. Möbius, R. Tozawa, M.C. Wehr, F. Wieland, S. Ishibashi, and K.A. Nave. 2005. High cholesterol level is essential for myelin membrane growth. Nature Neuroscience, 8(4):468–475.

[17] J. Tong, P.P. Borbat, J.H. Freed, and Y.K. Shin. 2009. A scissors mechanism for stimulation of SNAREmediated lipid mixing by cholesterol. Proceedings of the National Academy of Sciences, 106(13):5141.

[18] R.S. Tilvis, J.N. Valvanne, T.E. Strandberg and T.A. Miettinen. 2011. Prognostic significance of serum cholesterol, lathosterol, and sitosterol in old age; a 17-year population study. Annals of Medicine, Early Online, 1-10.

[19] L.R.Wagstaff, M.W. Mitton, B.M. Arvik, and P.M. Doraiswamy. 2003. Statin-associated memory loss: analysis of 60 case reports and review of the literature. Pharmacotherapy, 23(7):871–880.

[20] K.A. Weant and K.M. Smith. 2005. The Role of Coenzyme Q10 in Heart Failure. Ann Pharmacother. Sep; 39(9), 1522-6.

[21] G. Wainwright, L. Mascitelli, and M.R. Goldstein. 2009. Cholesterol-lowering therapy and cell membranes: stable plaque at the expense of unstable membranes? Arch Med Sci, 5:3.

[22] L. Zhuang, F. Jing, and X.Y. Zhu. 2006. Movie review mining and summarization. In Proceedings of CIKM, 43–50. ACM.

[23] Q. Zeng, S. Kogan, N. Ash, RA Greenes, and AA Boxwala. 2002. Characteristics of consumer terminology for health information retrieval. Methods of Information in Medicine, 41(4):289–298.

[24] P.E. Ziajka and T. Wehmeier. 1998. Peripheral neuropathy and lipid-lowering therapy. Southern Medical Journal, 91(7):667.

 ISBN: 978-1-61208-162-5

FATS: A Framework for Annotation of Travel Blogs Based on Subjectivity

Inmaculada Álvarez de Mon y Rego
Ingenieria Técnica de Telecomunicación. UPM
Lingüistica aplicada a la ciencia y la tecnología
Madrid, Spain
ialvarez@euitt.upm.es

Liliana Ibeth Barbosa Santillán
University of Guadalajara
Tecnologías de la Información
Guadalajara, México
ibarbosa@cucea.udg.mx

***Abstract*—This paper describes a framework for annotation on travel blogs based on subjectivity (FATS). The framework has the capability to auto-annotate -sentence by sentence- sections from blogs (posts) about travelling in the Spanish language. FATS is used in this experiment to annotate components from travel blogs in order to create a corpus of 300 annotated posts. Each subjective element in a sentence is annotated as positive or negative as appropriate. Currently correct annotations add up to about 95 per cent in our subset of the travel domain. By means of an iterative process of annotation we can create a subjectively annotated domain specific corpus.**

Keywords-Annotation; Subjectivity; Blogosphera; Spanish Language.

I. INTRODUCTION

The Social Web [1] has enabled humans to express and share their opinions and concerns to the World. One of the tools to share data and information within the net is a Blog. According to the WordPress [2] online dictionary, *a blog, or weblog,* is an online journal, diary, or serial published by a person or group of people.

Currently, there are several platforms such as Blogger APPis of Google [3] or APPis of MyBlogLog of Yahoo [4] that offer blogs classified according to several taxonomies including classes such as personal, business, non-profits, politics, etc [2]. Travel blogs belong to the personal class.

In this work, the authors take into consideration the following premises:

- According to [2] the type of blog, those studied belong to the personal class.
- The Blogs of study belong to the travel domain in Spanish language.
- The writing process used by bloggers to express their ideas depends on age, context, time, culture and geographic place.
- Despite the availability of good practices on the web to preserve the quality of writing, for instance FS250062 [5], most bloggers express their opinions in very heterogeneous ways.
- The blogs have several components including title, banner, tagboard, links, archives and posts.
- The posts of each blog keep their chronological order.
- Some of the elements of posts are positive or negative or both.

Considering all this, the scientific hypothesis of our research is to auto-annotate the bloggers' expressions based on their subjectivity with at least 90% recall and precision for posts. The methodology used in this research is top-down in contrast to the Folksonomies methodology [6] where the main aim is to annotate collaboratively the social web.

This research deals with corpora in two dimensions: linguistic and technical. The linguistic dimension refers to the selection of sentences and their elements of the posts. The technical dimension encodes, builds the meta-model for annotation and annotates the posts.

Linguistic dimension:

- The sentence selection is based on finding blogs with two main components: 1) the title, and 2) the first post.

Technical dimension:

- The encoding is based on the standard ANSI/NISO Z39.19-2005 [7] to represent parts of a sentence and, in addition, to add the mark P for positive and N for negative.
- Annotation is based on bracketing conventions of segments according to the recommendations for the morphosyntactic annotation of corpora in tagging from lexical data (EAGLES96) [8].
- The meta-model for annotation is represented by eight patterns proposed in this research (see Section IV).

In order to analyze the bloggers' expressions, a subset of an ad-hoc corpus collected by [9] is used. This corpus consists of 10 thousand words extracted from Spanish blogs, sampled from a comprehensive range of travel blogs.

The aim of this work is to automatically recognize the positive or negative elements of the posts and annotate them.

The remainder of the paper is organized as follows. Section II briefly discusses the related work. In Sections III and IV, the detailed functionalities of FATS are presented. Section V briefly evaluates the performance of the framework proposed. Finally, Section VI concludes this paper.

II. RELATED WORK

The related work dealt with takes into account three topics: subjectivity, lexicons and annotation.

 ISBN: 978-1-61208-162-5

A. Subjectivity

According to Wiebe [10], subjectivity is "the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs, and speculations".

Research work on subjectivity includes: Kanayama and Nasukawa [11] who built some domain-dependent polarity lexicons for Japanese language; Andreevskaia and Bergler [12] proposed various methods for learning subjectivity from WordNet. Esuli and Sebastiani [13] proposed a method for identifying both, the subjectivity and prior polarity of a word, also using WordNet. Wiebe and Mihalcea [14] were able to automatically identify whether a particular word was subjective or not, using a computer program. Kim and Hovy [15] used small seed sets and WordNet to identify sets of subjective adjectives and verbs, and Kobayashi et al. [16] identified domain-dependent sets of subjective expressions. This work is based on the subjectivity classification for the previous research work.

B. Lexicons

According to [17], a general definition of a lexicon is "the vocabulary of a language that contains all the words or LEXEMES in the language". A more specific one suitable for our domain dependent research is "Word stores that are primarily consulted for the reason of information retrieval are referred to as "dictionaries". By contrast, word-stores that constitute a component within a natural language processing system are called a lexicon [18], the LEXICON is understood broadly as a finite list of stored forms and the possibilities for combining them".

Previous research has focused on the creation of lexicons in English such as: Higashinaka [19] who used a set of dialogues to build her own lexicon. Lexicons are also available as linguistic resources in the Internet, some examples are SentiWordnet [20], NTU Sentiment Dictionary [21], Opinion Finders subjectivity Lexicon [22], etc. However our research relies on a controlled vocabulary that tries to eliminate noise by providing a list of preferred and non-preferred terms and a domain-semantic structure in Spanish. Therefore, the lexicon built by [9] was taken as reference. The process for creating this lexicon was the following: a) key terms used in valorative sentences are extracted and b) words or groups of words with positive or negative sentiment are selected for the lexicon. The final classes were nouns, adjectives, diminutives, prefixes, verbs, adverbs, interjections, and idioms taken from the language sample.

C. Annotation

Although there are many Spanish Corpora involving grammatical analysis or annotation of Spanish texts (e.g., Atwell [23]) and a significant number of software libraries developed at universities to annotate texts (Exmaralda [24] and MMax 2 [25] tools). However, no research has been found on annotating each component of a post -as positive or negative- as appropriate. The closest results were found to be by [23], [24], [25] and their results evaluated only part of speech tagging. Our proposal is based on the subjectively annotation of posts of blogs supported by a reference-subjectivity lexicon.

The Corpus of blogs used for this research consists of 10 thousand words of Spanish written blogs, sampled from a comprehensive range of blog within the travel domain. For this research annotation is the process of attaching subjectivity information to the posts which can be opinions, evaluations, emotions and beliefs. It consists of two main steps: identifying elements on the post, and attaching polarity information to these elements.

The process for annotation in our research is aligned with the automatic annotation of three linguistic levels: morphosyntactic, syntactic and semantic.

- At the morphosyntactic level, a word is divided into its root and suffixes.
- At the syntactic level, the lexico-syntactic representation of nouns, adjectives, verbs and adverbs (positive or negative or both) -all of them based on the lexicon proposed by [9]- are mapped according to the eight patterns proposed in Section IV.
- At the semantic level, the eight patterns (see Section IV) link the relationships between each word in each sentence of the posts.

III. Framework of FATS

The general architecture of FATS consists in a series of components performing sequential transformations on an input blog. The architecture is structured into four layers: Data Source, Matching Components, Analysis Components and Result Components, as shown in Fig. 1.

- *Data Source:* this layer contains the basic elements of blogs (see Fig. 2) required by the FATS.
- *Matching Components*: this second layer contains the main matching engineering functionality carried out through the analysis of components. Additionally, other components such as selector, split up, match, patterns and blogger-opinion are shown.
- *Analysis Components:* this middle layer contains the main linguistic levels of analysis in the blog: morphosyntactic, syntactic and semantic as explained in the introduction section.
- *Result Components*: the last layer contains the posts subjectively annotated of the blogs.

The relationships among the different components of the framework of FATS are explained below:

First of all, blogs are collected in digitized documents from the WWW throughout Heritrix [26] as an example the authors made a job limited by scope and frontier. Next, they are taken to the matching components where the blogs are structured and modelled on separated components such as

 ISBN: 978-1-61208-162-5

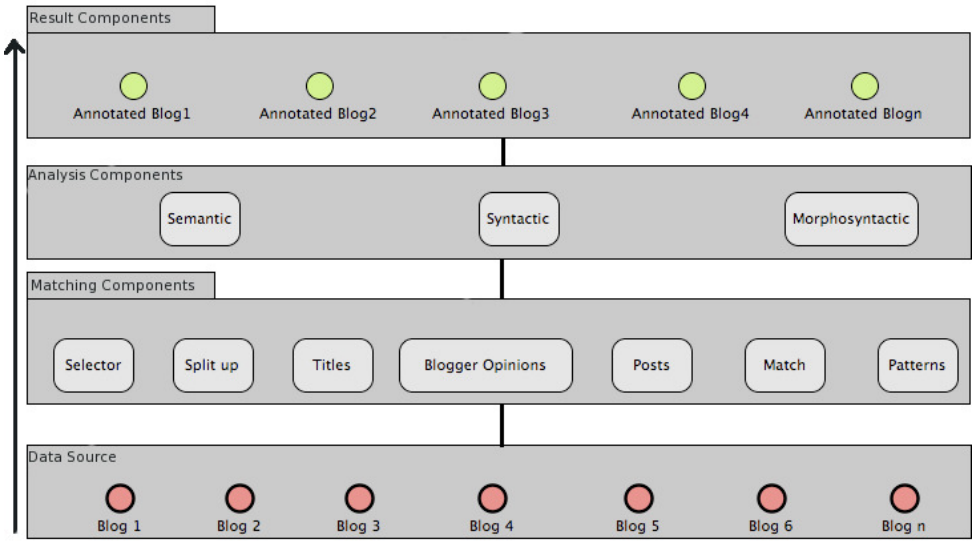


Figure 1. The general architecture of FATS.

titles and posts. Then, the selector component takes the first post and matches it with the related patterns (see Fig. 3); we do not match with the titles yet. In this stage, the analysis is conducted by the syntactic, morphosyntactic and semantic components. The syntactic analysis is done by mapping onto the lexico-syntactic representation of nouns, adjectives, verbs and adverbs (positive or negative or both) all of them based on the lexicon proposed by [9]. In the next stage, the morphological component processes a word into its smallest meaningful components, or morphemes. This is done by dividing a word into its root and suffixes. Subsequently, the semantic component establishes the relationships between each word in each sentence. Finally, the posts are placed in bags, which in turn have the annotated posts of blogs that determine the polarity of subjective expressions of the terms included in the posts.

IV. PATTERNS

For this research, we created eight different patterns as shown in Fig. 3, based on the most common structure type. Each pattern describes how the experiment interacts with the proposed framework (see algorithm 1) to achieve a new section of annotated posts. The 8 specialized patterns are represented by the following equations:

$$\forall x, y.Sentence(x, y) \rightarrow Subject(x) \sqcap Predicate(y). \quad (1)$$

$$P1 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \\ \sqcap Noun(y) \sqcap Pp(y) \sqcap Art(y) \sqcap Noun(y) \end{cases} \quad (2)$$

$$P2 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \\ \sqcap Noun(x) \sqcap Adj(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \\ \sqcap Noun(y) \sqcap Adj(y) \sqcap Pp(y) \\ \sqcap Art(y) \sqcap Noun(y) \sqcap Adj(y) \end{cases} \quad (3)$$

$$P3 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \\ \sqcap Art(y) \sqcap Noun(y) \end{cases} \quad (4)$$

$$P4 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \\ \sqcap Pp(y) \sqcap Art(y) \sqcap Noun(y) \end{cases} \quad (5)$$

$$P5 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \\ \sqcap Pp(y) \sqcap Noun(y) \end{cases} \quad (6)$$

$$P6 \doteq \begin{cases} \forall x.Subject(x) \rightarrow \exists Art(x) \sqcap Noun(x) \\ \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Pp(y) \\ \sqcap Noun(y) \end{cases} \quad (7)$$

$$P7 \doteq \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \sqcap Noun(y) \quad (8)$$

$$P8 \doteq \begin{cases} \forall y.Predicate(y) \rightarrow \exists Verb(y) \sqcap Art(y) \\ \sqcap Noun(y) \sqcap Pp(y) \sqcap Art(y) \sqcap N(y) \end{cases} \quad (9)$$

Some of the experiments are shown in Table 1 with 8 different subjective tagging schemes mentioned as follows:

 ISBN: 978-1-61208-162-5

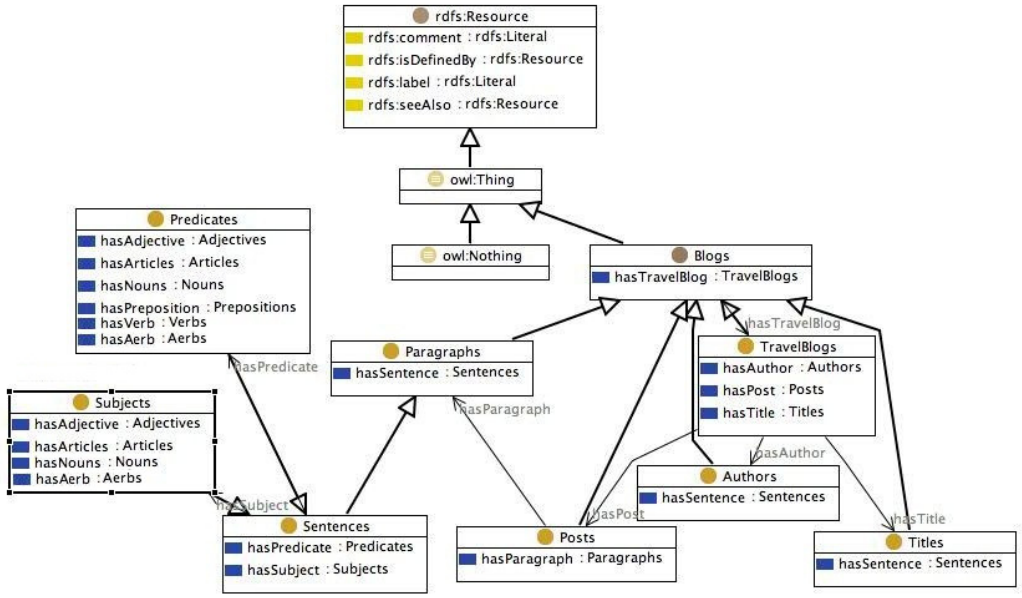


Figure 2. The structure of a blog

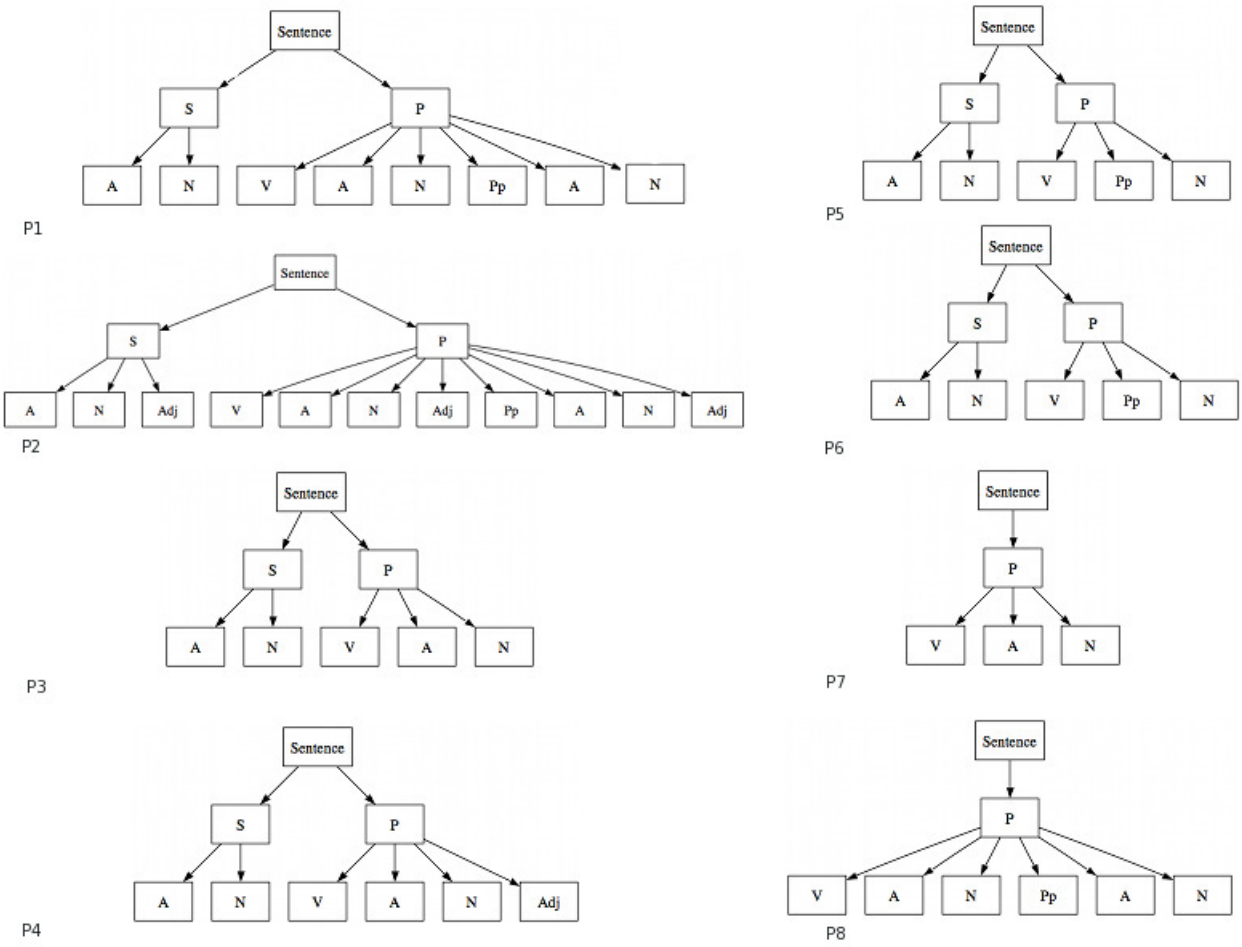


Figure 3. The 8 specialized patterns.

Table I
SOME OF THE EXPERIMENTS USING FATS

Blog	Post	Pattern	Annotated Post
1	Los dueños no-eran la alegría de la huerta. "The owners were not the joy of vegetable garden".	P1	s[nc[d[los], NP[dueños]], vc[VN[noeran], nc[d[la], NP[alegría], pp[p[de], nc[d[la], huerta]]]]],
2	El viaje perfecto continua sin incidentes malos, con un aprovechamiento máximo. "The perfect trip goes on without any bad incidents, with maximum benefit".	P2	s[nc[d[el], na[NP[viaje], AP[perfecto]]], vc[VP[continua], nc[d[sin], na[NP[incidentes], AN[malos]]], pp[p[con], nc[d[un], na[NP[aprovechamiento], AP[máximo]]]]]],
3	Los italianos son los abiertos. "Italians are open".	P3	s[nc[d[los], NP[italianos]], vc[VP[son], nc[d[los], NP[abiertos]]]],
4	El ambiente esta a la perfección. "The environment is perfect".	P4	s[nc[d[el], NP[ambiente]], vc[VP[esta], pp[p[a], nc[d[la], NP[perfección]]]]],
5	La verdad estabamos muy cansados. "Actually we were very tired."	P5	s[nc[d[la], NP[verdad]], vc[VP[estabamos], pp[p[muy], nc[NN[cansados]]]]],
6	Otra vez estamos super-cansados. "Again we are super tired."	P6	s[nc[d[otra], NP[vez]], vc[VP[estamos], pp[p[super], nc[NN[cansados]]]]],
7	Fuimos a Edimburgo. "We went to Edinburgh."	P7	s[vc[VP[fuimos], pp[p[a], nc[NP[edimburgo]]]]],
8	Tomamos un bus pero nos equivocamos. "We took a bus but we were wrong."	P8	s[vc[VP[tomamos], nc[d[un], NP[bus], pp[p[pero], nc[d[nos], NN[equivocamos]]]]]],

NP (noun positive), NN (noun negative), AP (adjective positive), AN (adjective negative), VP (verb positive), VN (verb negative), AdP (adverb positive), AdN (adverb negative). The part of speech tagging schemes are: s (subject), nc (noun complement), na (noun adjective), d (article), vc (verb complement), pp and p (preposition). Also, Table 1 presents an example of a post annotated using FATS. In this case the annotation scheme [8] was used to tag the posts elements. The bracketing of each element in the sentence involves the delimitation with square brackets.

V. PERFORMANCE RESULTS

In order to evaluate the quality of the structural markup, we calculated the recall and precision rates produced by FATS, as shown in formulas (10) and (11), where RW means Retrieved Words.

$$precision = \frac{|\{\text{relevant}(NP \cup AP \cup AdP \cup VP)\} \cap \{\text{RW}\}|}{|\{\text{RW}\}|} \quad (10)$$

$$recall = \frac{|\{\text{relevant}(NP \cup AP \cup AdP \cup VP)\} \cap \{\text{RW}\}|}{|\{\text{relevant (NP} \cup AP \cup AdP \cup VP)\}|} \quad (11)$$

The first task is to analyze whether each post is grammatically and semantically correct or not as shown in algorithm 1. If the post is incorrect grammatically/semantically, FATS ends the task of analysis and cannot continue with the process of annotation. However, anytime a post is both grammatically and semantically correct FATS produces a YES answer, at the same time comparing each element with the proposed patterns (see Section IV). The results are shown in Table 2 where a collection of 180 sentences of the 100 posts are tagged.

Algorithm 1 Annotating Posts of Blogs

```
1:  procedure APB(Posts,NumberofPosts)
2:      for i ← 1, NumberofPosts do
3:          for j ← 1, NumberofSentences(i) do
4:              if sentence(j) = pattern then
5:                  for k ← 1, NumberofElements do
6:                      element(k) ← syntactic(element(k));
7:                      element(k) ← mofpho(element(k));
8:                      element(k) ← semantic(element(k));
9:                      element(k) ← annotate(element(k));
10:                 end for
11:             end if
12:         end for
13:     end for
14: end procedure
```

Table II
RECALL AND PRECISION FOR THE POST

	NP	AP	VP	AdP
Recall	.93	.93	.93	.93
Precision	.93	.94	.94	.85

VI. CONCLUSIONS AND FUTURE WORKS

This paper presented FATS, a framework for tagging posts of blogs by using a table of symbols of subjectivity, a lexicon of reference [9], and eight patterns of analysis. Our framework automatically builds a subjectively annotated domain specific corpus by analyzing morphosyntactic, syntactic and semantic levels of posts with at least 90% recall and precision. The corpus resulting from this research can be used as it is for other applications because it is easy to integrate in other processes. We can see as shown in Fig. 4 that the accuracy of FATS in the tagging of NP, AP and VP is higher than in the tagging of AdP. However, the difference

 ISBN: 978-1-61208-162-5

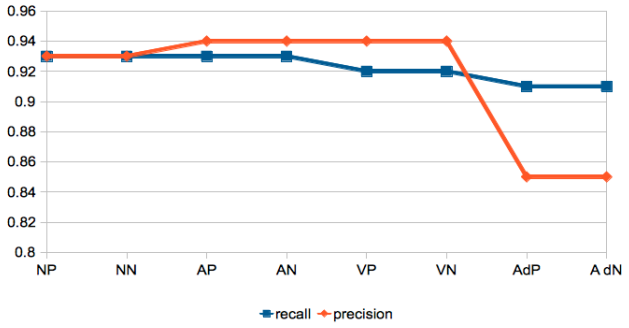


Figure 4. Precision for the posts.

is negligible and thus, it does not represent a limit for the present research. Future work will be done with ontologies to enrich and improve FATS.

ACKNOWLEDGMENT.

We are grateful to the Sciences Research Council (CONACYT) and COECYTJAL for funding this research project.

REFERENCES

[1] T. B.-L. Jim Hendler, "From the semantic web to social machines: A research challenge for ai on the world wide web," *Artif. Intell*, vol. 2, no. 174, pp. 156–161, 2010.

[2] http://en.wordpress.com/types-of-blogs/ (Accesed: June 2011).

[3] http://code.google.com/apis/blogger/ (Accesed: June 2011).

[4] http://mybloglog.com (Accesed: June 2011).

[5] T. S. A. in the UK, "Accessibility guidelines for written resources," *www.scoutbase.org.uk*, 2011.

[6] M. Muller-Prove, "Taxonomien und folksonomien tagging als neues hci-element (in german)," *I-com*, vol. 6, no. 1, pp. 14–18, 2007.

[7] NISO, "Guidelines for the construction, format, and management of monolingual controlled vocabularies." *ANSI/NISO Z39.19-2005 Bethesda, MD: National Information Standards Organization.*, 2005.

[8] http://www.ilc.cnr.it/EAGLES/annotate/annotate.html (Accesed: June 2011).

[9] M. R. Villarreal, "Corpus de blogs de viajes: análisis lingüístico para el reconocimiento de la valoración de la información. (in spanish)," *Proyecto Fin de Carrera. E. U. I. T. de Telecomunicación. Universidad Politécnica de Madrid.*, 2009.

[10] J. Wiebe, "Tracking point of view in narrative," *Proceedings of the NLPKE*, no. 174, 2008.

[11] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis." *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 353–363, 2006.

[12] A. Andreevskaia and S. Bergler, "Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses," *In Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, 2006.

[13] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," *In Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 193–200, 2006.

[14] J. Wiebe and R. Mihalcea, "Word sense and subjectivity," *In Proceedings of the 21st International Conference on Computational Linguistics*, 2006.

[15] S.-M. Kim and E. Hovy, "Automatic detection of opinion bearing words and sentences," *roceedings of the 2nd International Joint Conference on Natural Language Processing*, pp. 61–66, 2005.

[16] K. I. Y. M. K. T. Kobayashi, Nozomi and T. Fukushima, "Collecting evaluative expressions for opinion extraction," *Proceedings of the First International Joint Conference on Natural Language Processing*, 2004.

[17] SWANN, "Dictionary of socio linguistics," *The University of Alabama Press*, 2004.

[18] W. d. G. Jrgen Handke, *The structure of the lexicon: human versus machine*, 1995.

[19] R. Higashinaka, M. Walker, and R. Prasad, "Learning to generate naturalistic utterances using reviews in spoken dialogue systems," *ACM Transactions on Speech and Language Processing (TSLP)*, 2007.

[20] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06*, 2006, pp. 417–422.

[21] http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp (Accesed: June 2011).

[22] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Opinionfinder: a system for subjectivity analysis," in *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics, 2005, pp. 34–35.

[23] J. Kuhn, "Parsing word-aligned parallel corpora in a grammar induction context," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ser. ParaText '05. Association for Computational Linguistics, 2005, pp. 17–24.

[24] T. Schmidt, "Creating and working with spoken language corpora in exmaralda," in *LULCL II: Lesser Used Languages and Computer Linguistics II*, 2009, pp. 151–164.

[25] C. Müller, "Representing and accessing multi-level annotations in mmax2," in *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 73–76.

[26] http://crawler.archive.org/index.html (Accesed: June 2011).

 ISBN: 978-1-61208-162-5

The Migraine Radar - A Medical Study Analyzing Twitter Messages?

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

***Abstract*—This paper discusses the work in progress of the "Migraine Radar" project. The purpose of the project is to validate or disprove the assumed correlation between migraine attacks and weather conditions, especially weather changes. There have been various medical studies on this topic, but the correlation could not be proved with sufficient statistical significance so far. Furthermore, the results of some of the studies are contradictory. For this study, data from the micro-blogging platform Twitter will be analyzed. Twitter messages ("tweets") announcing currently or recently happened migraine attacks are retrieved using the Twitter API (Search-API, REST-API - Standard APIs provided by Twitter to retrieve tweet and user data). Weather data from weather information services are linked to the tweets, using the location information from Twitter. For the German language area, the results will be compared with the results obtained from a set of migraine announcements collected with the help of a web form in the same period of time. First statistics indicate that the number of migraine attacks announced in Twitter exceeds the number of cases in former classical studies by far. The project offers a wide range of possibilities to analyze Twitter messages with regard to migraine attacks. Beside the main purpose, it is also possible to analyze the distribution of migraine attacks over the weekdays or over the seasons. Furthermore an investigation of the spatial distribution of migraine attacks is possible. Instead of weather data, other information can be linked to the migraine sample as well. One example could be air pollution data.**

***Index Terms*—migraine; trigger; Twitter; weather; text mining**

I. Introduction

About 10% of the population suffer from migraine. Several factors are assumed to trigger migraine attacks. Examples for these potential trigger factors are food, stress, hormonal disbalance and sleep irregularities [1]. Especially weather conditions or changes in weather conditions are supposed to cause migraine attacks, although a clear correlation could not be proved in several studies. Typical problems of traditional studies in this area are that only a small number of patients were involved and only patients coming from a small local area were considered. Furthermore some studies were restricted to a short period of time or only took very severe cases of migraine attacks [2] into account.

Using data from Web 2.0 platforms like Twitter perhaps can help to solve some of these problems, but it must be admitted that other difficulties are expected to occur. For example, it could be a problem that the data will contain more noise insofar as in some cases there will be ambiguities while identifying the migraine tweets, e.g., it might be difficult to separate migraine and normal headache cases.

Thus, beside the main purpose of the study, another aim is to find out whether it is possible to use data from social networks to carry out or support medical studies and to identify typical problems while doing so.

In Section II, some related work is introduced. Section III contains a brief description of the Migraine Radar project and explains the main input data and how it is organized for further analysis. In Section IV, the data analysis approach is discussed. A short conclusion completes the paper in Section V.

II. Related Work

Several studies concerning the trigger factors of migraine attacks have been performed over the last decades. As weather conditions and changes in weather conditions are among the factors mentioned most frequently when discussing possible trigger factors, many studies have set their focus on the investigation of this correlation.

A former study investigated the correlation between weather changes and both tension headache and migraine [3]. Another purpose of this study was the investigation of the weekday dependence of migraine attacks. Weather data from a local Institute of Meteorology was used to enrich the clinical data of the migraine patients. The authors found evidence of a weekday dependence. They also saw a correlation to a change in atmospheric pressure 1-3 days after the attack. Limitations were the small number of patients under investigation and the restriction to a small local area.

Another study to investigate many possible trigger factors was performed using patient data from a headache clinic [1]. A detailed headache evaluation was performed, but without adding weather information to the data. Concerning the correlation between weather and migraine attacks, the author found out that about 50% of the patients report attacks occasionally triggered by weather changes.

The approach of Prince et al. [4] was somehow different. The authors evaluated headache calendars provided by 77 migraine patients in a headache clinic. They first asked the patients whether they believe that weather is a trigger for migraine attacks for them. Afterwards, they linked weather

 ISBN: 978-1-61208-162-5

data from the the National Weather Service to the data. Three weather factors were considered. The result was, that about 50% of the patients were found to be sensitive to at least one weather factor, more patients thought they were sensitive but were not.

A lot of research work on the analysis of Web 2.0 platforms is currently going on. Twitter as a micro-blogging platform offers some advantages compared to others. It provides an almost real-time access to utterances of user's daily life as well as insights into their thoughts, opinions and sentiments. In [5], Bifet and Frank give a brief introduction to Twitter as a micro-blogging platform. They also discuss the access to data using the Twitter API and some possibilities in the area of sentiment knowledge discovery using Twitter. In [6], the authors discuss the possibillity of monitoring earthquake occurencies using data from Twitter.

First studies in the medical sector have already been carried out using Twitter messages. In [7], for example, the author analyzed 500 million tweets to investigate an influenza outbreak in the United States enabling him to forecast future influenza rates with high accuracy. In [8], the authors present an automated tool for tracking the prevalence of influenza-like illness using Twitter messages.

Also Google search query data was already used to track influenza epidemics; see [9].

III. The Migraine Radar

A. General Overview

The project "Migraine Radar" collects announcements of migraine attacks from Twitter messages and from a web form. The data are enriched with weather data. Afterwards, the data are analyzed in order to investigate the correlation between the number of evident migraine attacks and certain weather conditions or changes to the weather conditions.

B. Input Data

Several data sources are important in the Migraine Radar project:

- Short messages announcing migraine attacks in the micro-blogging platform Twitter are retrieved with a simple search for "migraine" / "migräne" for the English / the German language. Due to the fact that the Twitter limit of 150 requests per hour is not reached, all relevant tweets are recorded. In some cases the spatial location of the patient is part of the tweet (about 2% of the tweets), in 80% of the cases it can be retrieved from the position the Twitter user has stored in his profile. If no position is available, the tweets will be discarded. Usernames are eliminated in order to anonymize the tweets. For normalization purposes (see section IV-C), tweets containing other search terms are also retrieved.
- Migraine attacks can also be reported using a web form [10], which is available only for Germany at the moment. In this form patients can add some additional information such as age, gender and severity of the migraine attack, making some additional analyses possible later on.
- In order to link the location of the tweets or the entries in the web form to an actual city or town, spatial information provided by GeoNames [11] is used.
- Historical and latest weather data are taken from weather information systems such as the "Deutscher Wetterdienst" [12] for Germany or the Weather Underground [13] for the United States.

C. Data Organization

All raw data are stored in a database. As soon as the preprocessing and the preselection of the migraine announcements is finished, the data are organized in a multidimensional data model to provide an easy and fast access for further analysis.

IV. Data Analysis

The analysis of the data will be performed in several steps. Firstly, a clean sample of migraine attack announcements with reliable spatial information and a defined start day of the attack has to be obtained (see chapters IV-A and IV-B). Afterwards, correlations between weather variables (e.g., temperature, pressure, wind, etc.) or their changes and the number of migraine attacks will be analyzed. Therefore, a normalization of the tweets is necessary (see chapter IV-C). Finally an investigation of the systematic uncertainties of the study has to be carried out (see IV-E).

A. Preprocessing and Preselection of the Tweets

Typical examples of tweets retrieved are as follows:

- *"UGH MIGRAINE"*
- *"Last Saturday, I woke up with an unbelievable migraine!"*
- *"Migraine solution for fast relief <some URL>"*
- *"@<some user> Maybe you're getting a migraine!"*
- *"Migraine, why aren't you going away? It hurts!"*

There are two important filtering steps required in order to select the tweets relevant for analysis:

1) Advertising: Tweets with recommendations, for example for medicine for migraine treatment, are normally not announcing an actual migraine attack. In most cases, they provide a web link to the advertised medicine. In order to eliminate tweets containing advertising, all messages that include web links are discarded.
2) Answering tweets: In most cases answers to other tweets are no real migraine attack announcements. Accordingly all tweets with "@"-signs have to be discarded.

A manual investigation of a part of the tweets indicates that more than 90% of the filtered tweets are announcements of migraine attacks and that the acceptance of good tweets is not affected too much by these steps. Precision and recall will be calculated during the final analysis to confirm this assumption.

B. Selection and Evaluation of the Tweets

The tweets remaining after the preselection steps have to be regarded in more detail. For each tweet it has to be decided:

1) Whether the tweet really refers to a migraine attack of the Twitter user writing the message.

 ISBN: 978-1-61208-162-5

2) At which time the migraine attack started. The first day of the attack is the relevant information, a finer resolution in time will not be achievable in the majority of cases. Some tweets announce actual attacks (like *"UGH MIGRAINE"*), some indicate a defined start day in the past (*"Yesterday I got a horrible MIGRAINE"*). But also ambiguous cases are occurring. As migraine attacks sometimes last up to three days, in the last example from above (*"Migraine, why aren't you going away? It hurts!"*), it cannot be decided when the attack started. Consequently tweets like this have to be discarded as well.

In order to perform this step of the analysis, text mining techniques have to be applied. One possible approach is to identify patterns which indicate an actual migraine attack (e.g., *"I ... have/had ... migraine"*) or to identify typical words occurring in migraine tweets (a first look at the remaining tweets shows that announcements of migraine attacks often contain swearwords).

Also some special cases have to be considered (e.g., *"migraine"* in names, in song titels etc.).

C. Normalization

In order to find out whether an increase in the number of migraine tweets really indicates a rise of migraine attacks under certain weather conditions or whether it is due to a systematic fluctuation, e.g., a general rise of Twitter activities, the tweets have to be normalized. As the total number of tweets for a specific date and place is not available, the normalization has to be obtained from tweets retrieved by searching for some other keywords for each of the two languages. For the English language, such words are for example "coffee", "time", "money", "nature" and "day". For the German language, the translated keywords ("Kaffee", "Zeit", "Geld", "Natur", "Tag") are used. In order to keep the number of normalization tweets small, individual downscaling factors are adjusted in a way that the number of downscaling tweets still exceeds the one of the migraine tweets.

The preselection criteria used for the migraine tweets are also applied to the normalization tweets. Especially tweets with web links (often containing advertising) and answers to other users (detected by the presence of "@"-signs) will be discarded.

D. Preliminary Data Collection Statistics

Table I summarizes the numbers of tweets retrieved per language for a period of four weeks. It also shows the effect of the preprocessing steps (elimination of advertisements and answers to other Twitter users). A first look at the remaining tweets shows that about 60% of them are tweets really announcing a migraine attack. This allows a rough estimation of the total number of migraine attacks recorded in a period of one year. A much more detailed analysis of the effect of the preprocessing steps remains to be done within the final analysis. Also the quality of the geo-location data, depending on whether the information is available directly from the tweet or only from the user profile, has to be analyzed.

Selection Step	**Number of Tweets**		
	English	**German**	**Sum**
Raw Data	81,444	1,057	82,501
Position Available	65,477	819	66,296
Without Links	56,355	600	56,955
Without @-signs	39,082	333	39,415
Exp. migraine tweets (4w)	$\approx$ 24,000	$\approx$ 200	
Exp. migraine tweets (1y)	$\approx$ 300,000	$\approx$ 2,400	

TABLE I
DATA COLLECTION STATISTICS FOR A FOUR WEEK CRAWLING PERIOD

E. Additional Remarks on the Data Analysis

Several follow-up studies have to be conducted in order to evaluate the systematic uncertainties of a study based on messages from Twitter:

- All the selection steps, including the preselection, have to be reviewed manually and performance indicators (e.g., precision, recall) have to be calculated.
- For the data sample in German, results obtained using the Twitter messages have to be compared with the results based on the data collected via the web form. As the information content of these data is richer and more reliable many systematic checks are possible. For example, the web form data contains a subset of messages (entered as migraine attacks with aura) where the probability of real migraine attacks is very high.
- The normalization can be the crucial point in the analysis. It has to be investigated, for example, the error of an improper normalization caused by different probabilities of normalization tweets during the day (e.g., there will be more *"coffee"* tweets in the morning).

F. Possibilities for Further Studies Using the Collected Data

The data collected for this study offer numerous further possibilities for doing research. Examples for other opportunities are:

- Investigation of the weekday dependence of migraine attacks: it is stated by many sources that migraine attacks occur more frequently on Saturdays or Sundays, because some of the attacks are triggered by decreasing stress. Several studies, e.g., Osterman et al. [3] and Morrison [14] investigated the weekday dependence of migraine attacks. They came to different results and the statistical significance was low. Similarly, season dependency of migraine attacks can be analyzed.
- Some migraine tweets contain information which allows to determine the starting time of the migraine attack (for example *"Migraine since three hours ..."*). Filtering these migraine tweets allows a study of the daytime dependence of migraine attacks with a chance to reach a higher statistical significance than former studies.

 ISBN: 978-1-61208-162-5

- The regional distribution of migraine attacks (e.g., urban vs. rural regions) can be examined.

V. CONCLUSION

Modern Social Media platforms like Twitter or Facebook offer enormous possibilities for information retrieval using modern data analysis techniques like data- and text mining.

The purpose of this study is to investigate the possibilities of conducting medical studies using data from Web 2.0 platforms. A first look at the data collected from Twitter seems to be promising, especially the number of migraine attacks analyzed is much higher than in classical clinical studies. But, of course, the approach discussed in this paper suffers from several limitations. In a classic clinical study, doctors diagnose migraine and ascertain information concerning the migraine. Using data from Web 2.0 platforms, it is not possible to verify that migraine announcements are really made by migraine patients.

First results of the project are expected by the end of 2011.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Institute of Information Systems (iisys) in Hof for many helpful discussions and especially R. Göbel for his great efforts on foundation of the institute.

This work is supported by the German Federal Ministry of Education and Research (BMBF). The Institute of Information Systems is supported by the Foundation of Upper Franconia and by the State of Bavaria.

REFERENCES

[1] L. Kelman, *The triggers or precipitants of the acute migraine attack*, Cephalalgia, 2007, **27**, pp. 394-402.

[2] W. J. Becker, *Weather and migraine: Can so many patients be wrong?*, Cephalalgia, 2011, **4**, pp. 387-390.

[3] P. O. Osterman, K. G. Lövstrand, P. O. Lundberg, S. Lundquist, and C. Muhr, *Weekly Headache Periodicity and the Effect of Weather Changes on Headache*, Int. J. Biometeor. 1981, vol. 25, number 1, pp. 39-45.

[4] P. B. Prince, A. M. Rapoport, F. D. Sheftell, S. J. Tepper, and M. E. Bigal, *The Effect of Weather on Headache*, Headache, 2004 Jun; 44(6), pp. 596-602.

[5] A. Bifet and E. Frank, *Sentiment knowledge discovery in Twitter streaming data*, Proceedings of the 13th international conference on Discovery science (DS'10), B. Pfahringer, G. Holmes, A. Hoffmann (Eds.) 2010, Springer-Verlag Berlin, Heidelberg, pp. 1-15.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World wide web (WWW '10), 2010, ACM, New York, pp. 851-860.

[7] A. Culotta, *Detecting influenza outbreakes by analyzing Twitter messages*, www2.selu.edu/Academics/Faculty/aculotta/pubs/culotta10detecting.pdf, (accessed July 26, 2011).

[8] Vasileios Lampos, Tijl De Bie, and Nello Cristianini, *Flu Detector - Tracking Epidemics on Twitter*, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2010, Springer-Verlag Berlin, Heidelberg, pp. 599-602.

[9] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457**, 2009, pp. 1012-1014.

[10] Institute of Information Systems, *The migraine radar*, www.migraene-radar.de (accessed July 26, 2011).

[11] Marc Wick (Founder), *GeoNames*, www.geonames.org (accessed July 26, 2011).

[12] Das Bundesministerium für Verkehr, Bau und Stadtentwicklung, *Deutscher Wetterdienst*, www.dwd.de (accessed July 26, 2011).

[13] Weather Underground Inc., *Weather Underground*, www.wunderground.com (accessed July 26, 2011).

[14] Morrison, *Occupational stress in migraine - is weekend headache a myth or reality?*, Cephalalgia, 1990, **10**, pp. 189-193.

 ISBN: 978-1-61208-162-5

Selecting Data Mining Model for Web Advertising in Virtual Communities

Jerzy Surma
Faculty of Business Administration
Warsaw School of Economics
Warsaw, Poland
e-mail: jerzy.surma@gmail.com

Mariusz Łapczyński
Department of Marketing Research
Cracow University of Economics
Cracow, Poland
e-mail: lapczynm@uek.krakow.pl

***Abstract*—Marketing in web based social networks (virtual communities) became one of the crucial topics in contemporary business. There are three reasons why this area of marketing usage is interesting. Firstly, there are more and more users of social media in the Internet. Secondly, access to behavioral data of potential clients permits acquiring knowledge about their needs. Finally, there is a possibility of direct one-to-one marketing communication. In this article we will present a study concerning an advertising campaign in a popular social network. We will use data mining methods in order to adjust the message to different users, and we will also present a method of choosing an appropriate communication to a given user when a class imbalance problem occurs. The results are very promising and point out that there is a need for further studies in the area of using data mining in marketing.**

Keywords-data mining; social network; web advertising; marketing campaign management

I. INTRODUCTION

At present virtual communities are closely studied by marketers who try to determine the customer behavioral process of how a given product is purchased. Due to access to user information in social network portals it is possible to target marketing messages even in ordinary approaches, such as web banners (banner ad) campaigns. This form of online advertising entails embedding an advertisement into a web page, and the advertisement is constructed from an image. When viewers click on the banner, they are directed (click-through) to the website advertised in the banner. Banner based advertisement campaigns on social networks portals may be monitored in real-time and may be targeted in a comprehensive way depending on the viewers' interests. It is possible because virtual community users are identified by a unique login and leave both declarative (sex, age, education, etc.) and behavioral (invitation sent/received, comments, usage frequency, etc.) data. Access to behavioral data constitutes a particular competitive advantage of an online social network as compared to other web portals. In this research we would like to focus primarily on investigating the potential supremacy of behavioral data mining for marketing campaign management based on web banners. Secondly, we would like to select the most suitable data mining techniques for this specific problem.

The main research problem is to optimize a marketing banner ad campaign by targeting a proper user, and to maximize the response measure by the click-through rate (response rate). We performed an empirical evaluation based on a marketing ad campaign for a cosmetic company. The problem of the response rate analysis and marketing campaign optimization is widely described in data mining textbooks [1][2], and more recently in the context of online social networks [3].

This research implies a class imbalance problem (banner click-through to display rate) that is described in Section II. In Section III, we discuss classification techniques (data mining models) that were chosen for this study. It is a comparison of existing data mining tools combined with sampling techniques whose goal is to overcome the class imbalance problem. In Section IV, we present a series of empirical experiments for selecting the best data mining model. Finally, in Section V the paper concludes with a summary of the experiments results.

II. CLASS IMBALANCE PROBLEM

A class imbalance problem is related to a situation when the number of objects belonging to one class (one category of dependent variables) is evidently smaller than the number of objects belonging to the other class. This problem is especially important in response analysis, where the customer reaction (in this case a click on the banner) is significantly lower than the number of messages (displays). In relationship marketing it refers to churn models, acquisition of customers, and in other disciplines to fraud detection, medical diagnosis, etc. Generally, there are two main approaches [4] to dealing with this problem; one is based on changing the structure of a learning sample (sampling techniques), while the other one pertains to cost-sensitive algorithms.

In the case of a heavily imbalanced class proportion the use of one-class learning is recommended [5]. The problem results from the fact that gathering information about the other class is sometimes very difficult, or the nature of the domain is itself imbalanced. Building classifiers by using cases belonging to one class would succeed in some situations. Some authors distinguish [6] between cost-sensitive learning and the so-called ensemble classifiers, i.e., based on the bootstrap procedure (bagging, random forests). However, this approach can be included in cost-sensitive learning algorithms. They are based on the CART algorithm [7] (Classification and Regression Trees) and utilize misclassification costs and a priori probabilities, just as CART does. Despite the existence of many ways of

 ISBN: 978-1-61208-162-5

overcoming skewed data, authors decided to concentrate on combining sampling techniques with cost-sensitive learning.

A. *Sampling Techniques for Imbalanced Datasets*

Up-sampling (also referred to as over-sampling) consists in replication of cases belonging to the minority class. It can be done randomly, directly, or by gathering synthetic cases, e.g., with the SMOTE algorithm [8]. In this research authors decided to use random up-sampling where cases from the positive response category are randomly multiplied.

Down-sampling (also referred to as down-sizing or under-sampling) consists in reducing the number of cases belonging to the majority class. Sometimes the elimination of overrepresented cases concerns redundant examples [9] or is based on Tomek's links concept [10]. For the purpose of this analysis the authors applied random down-sampling to balance the data set. These two methods of modifying data structure can be applied separately (one-sided sampling technique) or can be combined (two-sided sampling technique). Both of them were employed for the purpose of this research.

B. *Cost Sensitive Learning*

Cost sensitive learning is the next approach that can help to overcome the class imbalance problem. The goal of that type of building classifiers is to increase the prediction accuracy of cases belonging to the given category. Researchers should assign a different cost to a different misclassification of objects. Ling and Sheng [11] distinguish two categories of cost-sensitive learning. One of them is a set of direct algorithms, such as ICET or cost-sensitive decision trees. The other one is called cost-sensitive meta-learning methods, and it includes MetaCost, CSC (CostSensitiveClassifier), ET (Empirical Thresholding), or the cost-sensitive naïve Bayes. The difference between these two methods of dealing with skewed data consists in how they introduce misclassification costs.

It is worth explaining misclassification costs by using the cost matrix presented in Table 1. For example, TN is an acronym for true negative, which means that an object belonging to the negative category was classified as negative. Since TN and TP refer to correct classifications, costs are assigned to FN and FP. Building classifiers for a dichotomous dependent variable very often require that researchers focus on the positive class, and therefore the cost for FN should be greater than the cost for FP.

TABLE I. EXAMPLE OF COST MATRIX FOR TWO CATEGORIES OF DEPENDENT VARIABLES

		Classified	
		True	***False***
Observed	**True**	**TP** true positive	**FN** false negative
	False	**FP** false positive	**TN** true negative

In other words, it is more important to reduce the misclassification error of the positive class. If a higher cost is assigned to FN, one pays attention to avoid classifying a positive object as a negative one. Elkan [12] emphasizes the fact that costs cannot be treated only in monetary terms.

III. CLASSIFICATION METHODS

The following data mining models widely used in marketing applications were selected for the evaluation: single classification tree (the CART algorithm), random forests (RF) and gradient tree boosting. All these methods can apply a misclassification cost and different a priori probabilities.

CART, which was developed by Breiman et al, is a recursive partitioning algorithm. It is used to build a classification tree if the dependent variable is nominal, and a regression tree if the dependent variable is continuous. The goal of this experiment is to predict the customers' response, which means that a classification model will be developed. To describe it briefly, a graphical model of a tree can be presented as a set of rules in the form of if-then statements. A visualization of a model is a significant advantage of that analytical approach from the marketing point of view. Prediction is an important task for marketing managers, but the knowledge of the interest area is crucial. Despite the fact that CART was introduced almost thirty years ago it has some important features, i.e., a priori probabilities and misclassification costs, which make it potentially useful in cost sensitive-learning.

RF is a set of classification or regression trees used for predictive tasks that was developed by Breiman [13]. It combines a number of classifiers, and each of them is built by using a different set of independent variables. At every stage of the tree building procedure of a single tree (at every node) a set of explanatory variables is randomly chosen. The number of selected variables is usually denoted by the letter m, while the number of all variables is denoted by the letter M. The best split of a node is based on these m ($m<M$) predictors. Every single tree is built to its maximum possible extent without pruning. In the final stage trees vote on an object's class. Random forests are built by using bootstrap samples of the learning sample, as a result of which they usually outperform classic algorithms such as CART or C4.5.

Gradient tree boosting is based on the well-known concept of boosting [14] developed by Friedman in 1999 [15][16]. In short, a decision tree tries to assign an object to the given class. After the first attempt of prediction the cases belonging to a poorly classified class (usually the minority class) are given greater weight. At the next step a classifier uses that weighted learning sample and once again assigns a greater weight to the cases that were not classified correctly. During this iterative procedure many trees are built, and the sample voting procedure is applied while deploying model-based testing. It means that predictions from a single decision tree are combined to obtain the best possible output. Each classifier is induced from a bootstrap sample that is randomly drawn from the whole learning sample.

 ISBN: 978-1-61208-162-5

IV. EMPIRICAL EVALUATION

The dataset used in this experiment was obtained from the marketing campaign for a cosmetic company that was launched in the virtual community in October 2010. This ad campaign was especially focused on young women. The virtual community that was under investigation has several millions active users and a functionality similar to Facebook, and is mainly limited to users from one of the European countries. Every member of this virtual community was described by 115 independent variables and by one binary dependent variable. The set of the 115 independent variables consists of 3 declarative variables (sex, age, education) and 112 behavioral variables divided into four main subsets: on-line activity, interactions with others users, expenses, games. During the experiments 150,000 users were randomly selected and a double leaderboard banner was displayed (in accordance with the Interactive Advertising Bureau industry standard for the online advertising industry). During the one-week campaign the web banner was seen by 81,584 users, and 207 users clicked through (the response rate of 0.25%). These data proportions are highly skewed because of the small number of positive response cases. Table 2 shows the structure of the learning samples used in this study. The dataset was primarily divided into the learning sample (30%) and the test sample (70%). In the next step, the learning sample was modified in four ways as is shown below, and the test sample consists of 57,098 cases for all the four approaches L1-L4.

TABLE II. STRUCTURE OF LEARNING SAMPLES

Learning sample types		Learning samples		
		Positive response category	*Non-response category*	*Total*
L1	unmodified learning sample	59 (0.24%)	24,427 (99.76%)	24,486
L2	random up-sampling	590 (2.36%)	24,427 (97.64%)	25,017
L3	random under-sampling	59 (10%)	531 (90%)	590
L4	random up-sampling and random under-sampling	177 (10%)	1,593 (90%)	1,770

Four learning samples combined with three analytical tools (CART, RF and boosted trees), different misclassification costs as well as a priori probabilities (see details in Table 3) deliver 48 models. To compare all models presented in that article the following metrics were used:

- Accuracy = (TP + TN) / (TP + FP + TN + FN)
- True negative rate (Acc^-) = TN / (TN + FP)
- True positive rate (Acc^+) = TP / (TP + FN)
- Response rate = TP / (TP + FP)
- Profit (see details in Table 4).

TABLE III. ANALYTICAL MODELS FOR EMPIRICAL EVALUATION

Model		Misclassification costs	A priori probabilities
M1	CART	equal	equal
M2	CART	equal	75-25
M3	CART	10-1	estimated from data
M4	CART	20-1	estimated from data
M5	RF	equal	equal
M6	RF	equal	75-25
M7	RF	10-1	estimated from data
M8	RF	20-1	estimated from data
M9	BT	equal	equal
M10	BT	equal	75-25
M11	BT	10-1	estimated from data
M12	BT	20-1	estimated from data

Legend: RF – random forests, BT – boosting trees

TABLE IV. REVENUE-COST TABLE

	Revenue	Cost	Profit
TP	100	0.1	99.9
TN	0	-0.1	0.1
FP	0	0.1	-0.1
FN	-100	-0.1	-99.9

Table 5 compares the performance of different algorithms according to monetary costs and benefits of an advertising campaign. It turned out that three out of all the used learning samples, i.e., unmodified learning sample (L1), random under-sampling (L3) and two-sided sampling method (L4) made it possible to build effective classifiers. Random forests achieved a better performance than other algorithms. However, it cannot be replaced with a set of rules which are comprehensible for marketing managers. It is worth mentioning that the best CART models were based on L1, L3 and L4, while RF models with positive gains were based on L3 and L4. Models marked with "xxx" classified all instances as non-response. The best RF models have modified misclassification costs ratios and a priori probabilities, while the best CART models have modified a priori probabilities. In general, looking at positive gains one can notice that random under-sampling (L3) provides the best classifiers.

TABLE V. PERFORMANCE OF MODELS ACCORDING TO MONETARY PROFITS OF CAMPAIGN

Model		Learning sample			
		L1	*L2*	*L3*	*L4*
M1	CART	-1,464.7	-1,176.3	-2,184.3	-610.5
M2	CART	983.8	-1,176.3	1,855.5	996.0
M3	CART	-9,114.1	-1,176.3	-7,667.5	-3,780.1
M4	CART	-8,370.9	-1,176.3	-3,416.7	-1,770.8
M5	RF	-8,724.9	-7,594.6	-6,023.8	-6,450.9
M6	RF	-6,335.7	-4,740.9	2,892.6	1,114.1
M7	RF	-9,106.9	-2,021.6	2,177.1	4,119.6
M8	RF	xxx	-4,251.1	7,465.0	2,309.8
M9	BT	xxx	xxx	-9,107.9	-8,236.6
M10	BT	xxx	xxx	-9,106.9	-9,124.7
M11	BT	xxx	xxx	-9,114.1	xxx
M12	BT	xxx	xxx	-9,114.1	xxx

Legend: RF – random forests, BT – boosting trees

 ISBN: 978-1-61208-162-5

Tables 6 and 7 summarize performance metrics for learning samples L1 and L2. To compare differences between models the G-test at the 95% confidence interval was conducted. The results marked with an asterisk (*) signify lack of difference between the best results in the given column. As far as accuracy, true negative rate and response are concerned, RF outperforms other algorithms. However, CART models (M2 and M3) seem to be effective as well. One can hardly tell the difference between the unmodified learning sample and random up-sampling. It is important to note that CART models (M1-M4) deliver better results according to the true positive rate (Acc+).

TABLE VI. PERFORMANCE METRICS FOR UNMODIFIED LEARNING SAMPLE (L1)

Model	L1			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.512	0.512	0.446	0.002*
M2	0.429	0.429	0.561*	0.003*
M3	0.997*	0.999	0.000	0.000
M4	0.992	0.994	0.027	0.012
M5	0.996	0.998	0.014	0.022
M6	0.820	0.822	0.162	0.002*
M7	0.997*	1.000*	0.000	0.000
M8	xxx	xxx	xxx	xxx
M9	xxx	xxx	xxx	xxx
M10	xxx	xxx	xxx	xxx
M11	xxx	xxx	xxx	xxx
M12	xxx	xxx	xxx	xxx

TABLE VII. PERFORMANCE METRICS FOR RANDOM UP-SAMPLING (L2)

Model	L2			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.503	0.503	0.459*	0.002
M2	0.503	0.503	0.459*	0.002
M3	0.503	0.503	0.459*	0.002
M4	0.503	0.503	0.459*	0.002
M5	0.972*	0.975*	0.061	0.006*
M6	0.855	0.857	0.203	0.004*
M7	0.726	0.727	0.345	0.003*
M8	0.793	0.794	0.243	0.003*
M9	xxx	xxx	xxx	xxx
M10	xxx	xxx	xxx	xxx
M11	xxx	xxx	xxx	xxx
M12	xxx	xxx	xxx	xxx

Tables 8 and 9 display performance metrics for random under-sampling (L3) and a combination of up-sampling with under-sampling (L4). To compare the differences between models the G-test at 95% confidence interval was conducted, too. The best accuracy is provided by boosted trees models based on L3. As to the true negative rate (Acc-), it is hard to decide clearly which model and sampling method is superior. Random forests built on L3 with modified misclassification costs (M8) provide the highest true positive rate (Acc+). CART and RF deliver comparable results from the response point of view. It is hard to indicate which approach is the best.

TABLE VIII. PERFORMANCE METRICS FOR RANDOM UNDER-SAMPLING (L3)

Model	L3			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.694	0.695	0.351	0.003*
M2	0.348	0.348	0.622	0.002*
M3	0.949	0.951	0.068	0.004*
M4	0.761	0.762	0.284	0.003*
M5	0.865	0.867	0.155	0.003*
M6	0.352	0.351	0.655	0.003*
M7	0.411	0.411	0.608	0.003*
M8	0.122	0.120	0.899*	0.003*
M9	0.997*	1.000*	0.000	0.000
M10	0.997*	1.000*	0.000	0.000
M11	0.997*	0.999	0.000	0.000
M12	0.997*	0.999	0.000	0.000

TABLE IX. PERFORMANCE METRICS FOR RANDOM TWO-SIDED SAMPLING TECHNIQUE (L4)

Model	L4			
	Accuracy	*Acc-*	*Acc+*	*Response*
M1	0.657	0.658	0.419	0.003*
M2	0.430	0.430	0.561	0.003*
M3	0.729	0.730	0.284	0.003*
M4	0.538	0.538	0.426	0.002*
M5	0.915	0.917	0.122	0.004*
M6	0.493	0.493	0.541	0.003*
M7	0.389	0.388	0.682*	0.003*
M8	0.301	0.300	0.655*	0.002*
M9	0.986	0.989	0.034	0.008*
M10	0.996*	0.998*	0.000	0.000
M11	xxx	xxx	xxx	xxx
M12	xxx	xxx	xxx	xxx

In order to understand the results in a more comprehensive manner we applied a lift chart, which is a widely used graphical presentation of how the lift measure changes in population (see Figure 1). A lift measure is the ratio between a modeled response and a random response. The modeled response is provided by a statistical or data mining predictive model and is presented as a lift curve. The random response is sometimes called the base rate, and this is the response percentage in the whole population.

The denominator of a lift measure is presented as the baseline on the graph. The bigger the surface between the baseline and the lift curve, the better the model is. The X axis represents the percentage of the population in order of decreasing probability of belonging to the positive response class. On the Y axis there are cumulative lift values for every decile of population. Lift values greater than one mean that the model performs better than random targeting.

 ISBN: 978-1-61208-162-5

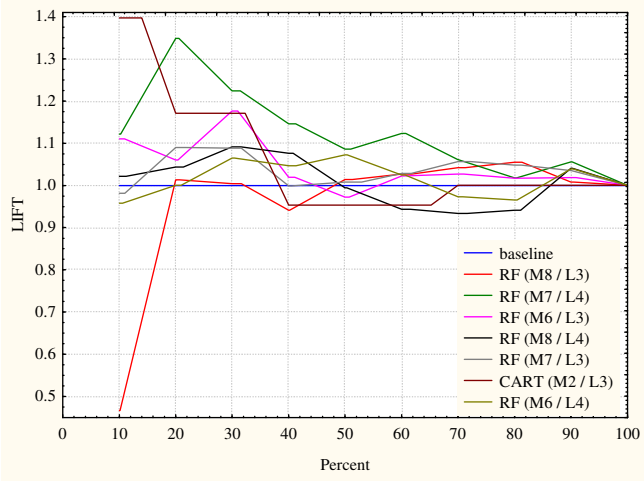


Figure 1. Lift chart for 6 best RF models and best CART model (profit > 1000)

The LIFT chart shows that the best results for the first decile are provided by the CART model with under-sampling, and the best results for the first two deciles are provided by the RF model with modified misclassification costs (10-1) based on L4. The line related to the best random forests model lies below the baseline, which means that cases with the highest predicted probability of belonging to the positive response category were incorrectly classified.

Additionally, we use the gain chart (see Figure 2), which is the second graphical tool that illustrates model performance. The percentage of the target population is shown on the X axis in descending order. The Y axis represents the cumulative percentage of target. The gain curve indicates the cumulative percentage for 1st class in the given percentage of the population, e.g., customer database. The gain chart confirms the interpretation of the LIFT chart. If one decides to use the CART model, 19.6% response rate can be achieved by showing a banner advertisement to 14% of website users with the highest predicted probability.

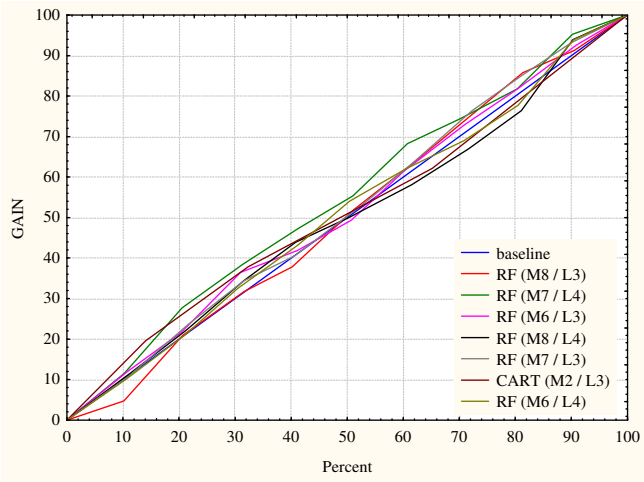


Figure 2. Gain chart for 6 best RF models and best CART model (profit > 1000)

The results of this study indicate that the best classifier can be obtained by combining under-sampling with cost-sensitive learning and random forests. The next best solution is to use the two-sided sampling method with cost-sensitive learning and RF. For the first decile of the test sample the CART model outperforms random forests. In general, the true positive rate and response were not satisfactory in such highly skewed data.

V. CONCLUSIONS AND FINAL REMARKS

The RF approach is clearly predominant but has at least one significant disadvantage, i.e., lack of clear model interpretation. This might be really problematic for marketers. On the other hand, the CART model has performed quite well and is easy to interpret. In this context one of the most astonishing discoveries is the set of variables established by the CART model. This model is completely based on the behavioral attributes with one exception, i.e., the age variable, which is of relatively low importance. In fact young women were the target group for this marketing campaign. We performed an additional experiment and displayed the advertisement directly to the target group. The received response rate (0.26%) was significantly lower than in the CART approach. Therefore, the standard segmentation approach might be augmented by an analysis of behavioral data in virtual communities.

Additionally, we should comment on the cost analysis results. We have found out that if the ratio between cost and revenue is lower than 0.0001 (in fact the cost of an ad banner display is normally significantly lower comparing to potential profits from the acquisition of new customers), it is better to send the web banner to all the available users. In this situation the cost of displays to FP users is covered by profits (the maximum number of TP hits). It is quite a reasonable approach because banner ads do not have the same bad impact as e-mail spam.

This specific context of web advertising in social networks should be investigated in the future research. An additional area for future research is to check if overcoming a class imbalance problem may be achieved by using predictors from RF variable importance ranking to build logit models. Treating random forests as a feature selection tool is a common practice.

REFERENCES

[1] R. Nisbet, J. Elder, and G. Miner, “Handbook of Statistical Analysis and Data Mining Applications,” Elsevier, Amsterdam, 2009.

[2] S. Chiu and D. Tavella, “Data mining and market intelligence for optimal marketing returns,” Elsevier, Amsterdam, 2008.

[3] J. Surma and A. Furmanek, “Improving marketing response by data mining in social network,” The 2nd International Workshop on Mining Social Networks for Decision Support, Odense, 2010.

[4] C. Chen, A. Liaw, and L. Breiman, “Using random forest to learn unbalanced data,” Technical Report 666, Statistics Department, University of California at Berkeley, 2004.

 ISBN: 978-1-61208-162-5

[5] B. Raskutti and A. Kowalczyk, "Extreme rebalancing for SVMs: a case study," SIGKDD Explorations , 2004.

[6] S. Hido and H. Kashima, "Roughly balanced bagging for imbalanced data," In SDM 2008, SIAM, 2008, pp. 143-152.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," Belmont, CA: Wadsworth International Group, 1984.

[8] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," ECML/PKDD, 2008.

[9] M. Kubat and S. Matwin, "Adressing the curse of imbalanced training sets: one-sided selection," Proc. 14th Intl. Conf. on Machine Learning", 1997, pp. 179–186.

[10] I. Tomek, "Two modifications of CNN. IEEE Trans. on Systems," Man and Cybernetics 6, 1976, pp. 769–772.

[11] C. X. Ling and V. S. Sheng, "Cost-Sensitive Learning and the Class Imbalance Problem", Encyclopedia of Machine Learning. C. Sammut (Ed.). Springer Verlag, Berlin, 2008.

[12] C. Elkan, "The Foundations of Cost-Sensitive Learning," In Proc. of the Seventeenth International Joint Conference of Artificial Intelligence, Seattle, Washington, Morgan Kaufmann, 2001, pp. 973-978.

[13] L. Breiman, "Random Forests," Machine Learning, 45, 5–32, Kluwer Academic Publishers, 2001, pp. 5-32.

[14] Y. Freund and R. Shapire, "Experiments with a new boosting algorithm," Machine Learning, Proc. of the Thirteenth International Conference 1996, pp. 148-156.

[15] J. H. Friedman, "Greedy Function Approximation: a Gradient Boosting Machine," Technical Report, Department of Statistics, Stanford University, 1999.

[16] J. H. Friedman, "Stochastic Gradient Boosting," Technical Report, Department of Statistics, Stanford University, 1999.

 ISBN: 978-1-61208-162-5

Large-Scale Association Rule Discovery from Heterogeneous Databases with Missing Values using Genetic Network Programming.

Eloy Gonzales*, Takafumi Nakanishi* and 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, takafumi, zettsu}@nict.go.jp

***Abstract*—Association Rule Mining is an important data mining task and it has been studied from different perspectives. Recently multi-relational rule mining algorithms have been developed due to many real-world applications. However, current work has generally assumed that all the needed data to build an accurate model resides in a single database. Many practical settings, however, require the combination of tuples from multiple databases to obtain enough information to build appropriate models for extracting association rules. Such databases are often autonomous and heterogeneous in their schemes and data. In this paper, a method for association rule mining from large, heterogeneous and incomplete databases is proposed using an evolutionary method named Genetic Network Programming (GNP). Some other association rule mining methods can not handle incomplete data directly. GNP uses direct graph structure and is able to extract rules without generating frequent itemsets. The performance of the method is evaluated using real scientific heterogeneous databases with a high rate of missing data.**

Keywords-Association rule mining; heterogeneous databases; missing values; evolutionary computation.

I. INTRODUCTION

Data mining has emerged as an important area mainly due to the rapid growth of the size and number of databases in a variety of scientific and commercial domains. It had generated a great need for discovering knowledge hidden in large and heterogeneous databases. Thus, recently, data mining techniques focus on finding novel and useful patterns or rules from this kind of databases. Traditionally, data mining algorithms have focused on relational databases and assumed that all relevant information for building a model is present within a single database. Moreover, it is also assumed that the records in the databases are always complete. However, in today's real scenarios, the sources of information for effective data mining algorithms rely on a large number of diverse, heterogeneous, incomplete but interrelated data sources. That implies the combination of records from multiple databases to obtain enough information to build an accurate data mining model. One of the most important tasks in data mining is association rule mining, which is the process of identifying frequent patterns from a dataset that usually require some minimum support and minimum confidence. Then, they allow the construction of association rules which portray the patterns as predictive relationships between particular attribute values. During the last decade, many promising techniques for association rule mining [1][2] have been proposed which achieved effective performances. However, none of them handle incomplete databases. Most of the techniques either eliminate the missing values or replace them with an average or mean value. Nevertheless, it is not possible for all the types of datasets to fill with mean values or frequency, such as the combination of several heterogeneous and diverse databases. Therefore, new algorithms for extraction of interesting association rules directly from incomplete databases are necessary.

In this paper, a method for extracting general association rules from databases with missing values is proposed using an evolutionary optimization technique named Genetic Network Programming (GNP). The *missing completely at random* is the missing data induction mechanism considered because the missing data in the attributes of databases are independent on either the observed or the missing data. [3]. There have been some proposals of association rule mining using GNP [4][5]. Class association rules from incomplete datasets using GNP have been proposed [6][7], however these approaches are only effective in mining class association rules whose consequent parts are restricted within a class label. In this work, an extended method for mining general association rules from incomplete datasets is presented, which uses the *cosine measure* to evaluate the correlation of rules.

The following sections of this paper are organized as follows: In Section II, the concepts and explanations of general association rules are presented, the explanation of incomplete databases is introduced in Section III, the outline of GNP is briefly reviewed in Section IV where also the method for rule extraction from incomplete databases is presented. Simulation results are described in Section V, and finally, conclusion and future work are given in Section VI.

 ISBN: 978-1-61208-162-5

II. Association Rules

In this section, the definition and properties of association rules are briefly reviewed. The following is a formal statement of the problem of mining association rules [8]. Let $I = \{A_1, A_2, \ldots A_l\}$ be a set of attributes. Let G be a set of transactions,where each transaction T is a set of attributes such that $T \subseteq I$. Associated with each transaction is a unique identifier whose set is called TID. A transaction T contains X, a set of some attributes in I, if $X \subseteq I$. 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. Both are called **itemsets**. In general, an itemset is a non-empty subset of I.

Each itemset has an associated measure of statistical significance called support. If the fraction of transactions containing X in G equals *t*, then $support(X) = t$. The rule $X \Rightarrow Y$ has a measure of its strength called confidence defined as the ratio of $support(X \cup Y)/support(X)$. This measure indicates the relative frequency of the rule, that is, the frequency with which the consequent is also fulfilled when the antecedent is fulfilled .

The support-confidence framework is the most widely used model for mining association rules. The algorithm works in two phases, first searching of frequent itemsets in a database and then extract all association rules meeting user-specified constraints such as minimum support and minimum confidence. However, this framework is not enough for extracting interesting association rules [9], therefore additional correlation measures such as lift, chi-squared, cosine, etc. are very useful and convenient to improve the quality of the extracted rules. In this paper, *cosine correlation measure* is used in addition to support-confidence framework because it ensures that only positive correlation rules are extracted [10].

Given two itemsets X and Y, the cosine measure [10] is defined as:

$$cosine(X,Y) = \frac{P(X \cup Y)}{\sqrt{P(X)\ P(Y)}} = \frac{supp(X \cup Y)}{\sqrt{supp(X)\ supp(Y)}} \quad (1)$$

where,

$P(X \cup Y)$ is the probability of taking X and Y.
$P(X)$ is the probability of taking X.
$P(Y)$ is the probability of taking Y.
$supp(X \cup Y)$ is the support of X and Y.
$supp(X)$ is the support of X.
$supp(Y)$ is the support of Y.

Cosine is a number between 0 and 1. This is due to the fact that both $P(X \cup Y) \leq P(X)$ and $P(X \cup Y) \leq P(Y)$ are satisfied. A value close to 1 indicates a positive correlation between X and Y. The total number of transactions N is not taken into account by the cosine measure. Cosine measure is **null-invariant** because its value is not influenced by **null-transactions**. A **null-transaction** is a transaction that does not contain any of the itemsets being examined. Null-invariance is an important property for measuring correlations in large databases especially in the case of missing values.

III. Association Rules within an incomplete database

Most of the conventional association rule mining algorithms assume that databases are complete. Generally, databases are pre-processed in order to eliminate missing values or to replace them with an average or other statistical measures because the main problem in such kind of datasets is the difficulty for calculation of measures such as $support$, $confidence$ and $cosine$.

Table I
Example of database with missing data

TID	A_1	A_2	A_3	A_4	A_5	A_6
1	1	1	1	1	1	0
2	0	1	m	0	1	1
3	1	1	m	1	1	1
4	0	0	1	m	0	1
5	1	0	0	m	0	1
6	0	0	m	1	1	0
7	1	m	1	1	1	1
8	1	m	0	m	0	1
9	0	m	m	1	1	1
10	0	1	1	0	0	0

Table I is an example of an incomplete database which contains missing values. A_i is an attribute in the database. Missing data is represented as *"m"*, a different value of 1 or 0.

Considering Table I, the measurements for association rules from incomplete databases are calculated as follows: In case of the rule $(A_1) \rightarrow (A_5) \wedge (A_6)$, tuple $TID = 3$ includes A_1, A_5 and A_6, but tuple $TID = 10$ does not include neither A_1, A_5 and A_6. Notice that tuple $TID = 3$ contains missing data, however all records ($N = 10$) in the database are available for calculation of the measurements because it is possible to judge whether each record satisfy the rule or not. Consequently the measurements of the rule are: $support((A_1) \rightarrow (A_5) \wedge (A_6)) = 2/10$ and $confidence((A_1) \rightarrow (A_5) \wedge (A_6)) = 2/5$ as usual.

In the case of rule $(A_2) \wedge (A_5) \rightarrow (A_6)$, it is clear that tuples $TID = 2$ and $TID = 3$ satisfy completely the rule. Tuple $TID = 1$ does not satisfy the rule because it does not include A_6, that is $A_6 = 0$, the same as tuples $TID = 4$, $TID = 5$, $TID = 6$, $TID = 8$ and $TID = 10$ which contain at least one attribute whose value is 0 and therefore they surely do not satisfy the rule. However, these tuples are available for calculating the measurements. On the other hand, it is not possible to judge whether tuples $TID = 7$ and $TID = 9$ satisfy the rule or not because of the missing

 ISBN: 978-1-61208-162-5

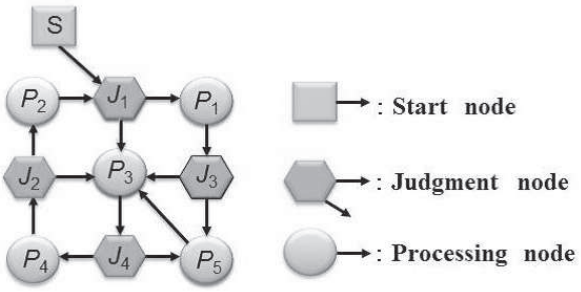


Figure 1. Basic structure of GNP

information of A_2; therefore, these tuples are omitted for the calculation of the measurements. Thus, the measurements of the rule are: $support((A_2) \wedge (A_5) \rightarrow (A_6)) = 2/8$ and $confidence((A_2) \wedge (A_5) \rightarrow (A_6)) = 2/3$.

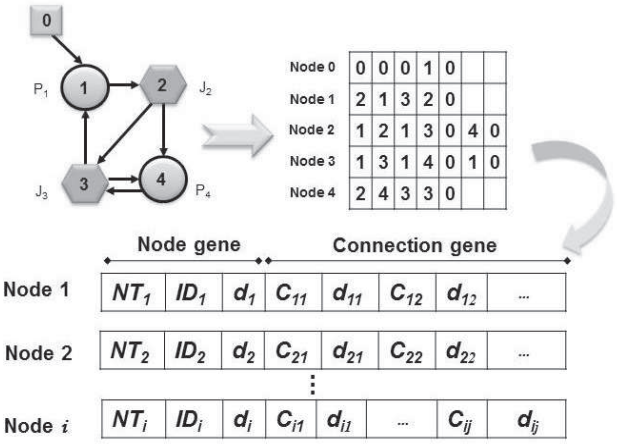
It is clear that the number of records N to be considered for the calculation of the measurements are different rule by rule. The available records for each rule are calculated according to the matching of the rule with the records. In other words, a record is included when it is ensured that it does not satisfy the rule (it contains any attribute with value 0) despite of it may contain missing values. Conversely, a record is excluded when it is not possible to judge if it satisfies the rule or not by missing values. Obviously, in the case of a complete database, i.e., with no missing data, N represent the total number of tuples in the database.

IV. Genetic Network Programming

Genetic Network Programming (GNP) is one of the evolutionary optimization algorithms, which evolves directed graph structures as solutions instead of strings (Genetic Algorithms) or trees (Genetic Programming) [11], [12], [13]. 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.

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



NT_i : node type (Start node=0; Judgment node=1; Processing node=2)
ID_i : identification number; d_i, d_{ij}: delay time; C_{ij} : connected node

Figure 2. Gene structure of GNP (node i)

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. All GNP-individuals in a population have the same number of nodes.

The characteristics of GNP are described as follows. (1) The judgment and processing nodes are repeatedly used in GNP, therefore the structure becomes compact and an efficient evolution of GNP is obtained. (2) Since the number of nodes is defined in advance, GNP can find the solutions of the problems without bloating, which can be sometimes found in Genetic Programming (GP). (3) Nodes that are not used at the current program execution will be used for future evolution. (4) GNP is able to cope with partially observable Markov processes. (5) The node transition in GNP individual is executed according to its node connections without any terminal nodes.

In the conventional GNP-based mining method, the attributes of the database correspond to the judgment nodes in GNP. Association rules are represented by the connections of nodes. Candidate rules are obtained by genetic operations. Rule extraction using GNP is done without identifying frequent itemsets used in Apriori-like methods [14]. Therefore, this method extracts important rules sufficient enough for user's purpose in a short time. The association rules extracted are 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

 ISBN: 978-1-61208-162-5

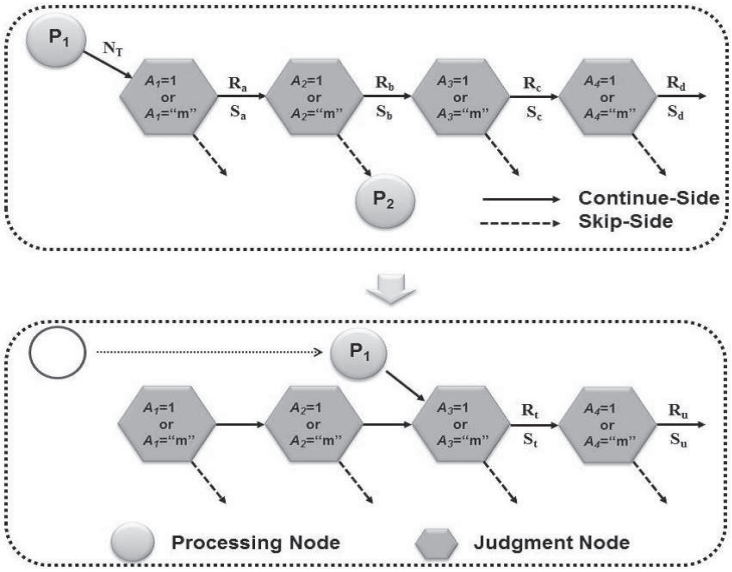


Figure 3. A connection of nodes in GNP for association rule mining with missing values

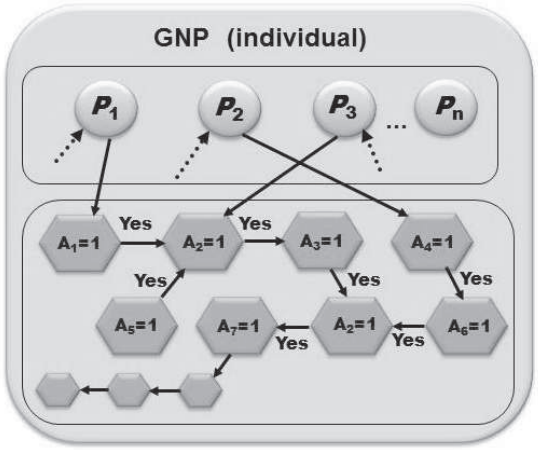


Figure 4. Basic structure of GNP for association rule mining

fitness value. GNP method has also advantages over other evolutionary methods such as Genetic Algorithms (GA) and Genetic Programming (GP). For GA-based methods [15], there are limitations in the number of association rules extracted because they are represented in individuals. In GP-base methods [16], an individual is usually represented by a tree with attribute values in the functions (e.g., logical, relational or mathematical operators) of the internal nodes. An individual's tree can grow in size and shape in a very dynamical way making it very difficult to understand for real applications.

A. GNP for rule extraction in an incomplete database

In this section, a general association rule mining method for incomplete databases is proposed using GNP. Let A_i be an attribute in an incomplete binary database and its value be 1, 0 or "m".

1) Rule Representation: Attributes and its values correspond to the functions of judgment nodes in GNP. Association rules are represented as the connections of nodes .

Fig. 3 shows a sample of the connection of nodes in GNP for association rule mining. P_1 is a processing node and is a starting point of association rules. "$A_1 = 1$", "$A_2 = 1$", "$A_3 = 1$" and "$A_4 = 1$" in Fig. 3 denote the functions of judgment nodes. Association rules are represented by the connections of these nodes, for example, $(A_1 = 1) \Rightarrow (A_2 = 1)$, $(A_1 = 1) \wedge (A_2 = 1) \Rightarrow (A_3 = 1)$, $(A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1) \Rightarrow (A_4 = 1)$ and $(A_1 = 1) \wedge (A_2 = 1) \Rightarrow (A_3 = 1) \wedge (A_4 = 1)$.

Judgment nodes in GNP are used to examine the attribute values of database tuples and processing nodes calculate the measurements of association rules. Judgment nodes determine the next node by a judgment result. Each judgment node has two connections Continue-side and Skip-side. The Continue-side of the judgment node is connected to another judgment node. Skip-side of the judgment node is connected to the next numbered processing node. If the attribute value is 1 or "m", then move to the Continue-side. If the attribute value is 0, then the transition goes for the Skip-side.

For example in Table I tuple $TID = 5$ satisfy $A_1 = 1$ and $A_2 = 0$, therefore a transition from P_1 to P_2 occurs in the upper side of Fig. 3

A basic structure of GNP-individual for association rule mining is shown in Fig. 4. In Fig. 4, the Skip-side of judgment nodes is abbreviated.

Each processing node has an inherent numeric order (P_1, P_2, ..., P_s) and is connected to a judgment node. Start node connects to P_1. For each judgment node, the examinations of attribute values start and in case to move to the Continue-side continuously, the connection is obligatorily transfered to the next processing node using the Skip-node when the maximum number of attributes ($MaxLength$) in the rule is reached.

When the examination of the attribute values of tuple $TID = 1$ from the starting point P_s ends, then GNP examines the next tuple $TID = 2$ from P_1 likewise. Therefore, all tuples in the database are examined.

2) Rule Measurements: In GNP the number of tuples moving to the Continue-side are counted up and they are used for calculation of the measurements In Fig. 3, R_a, R_b, R_c and R_d are the number of tuples moving to the Continue-side at each judgment node when the attribute value is only 1. On the other hand, S_a, S_b, S_c and S_d represent the number of tuples moving to the Continue-side at each judgment node when the attribute value is 1 or "m". Therefore, the number of available records (N_x) for calculation of the rule measurements is given by the following equation:

$$N_x = N_T - (S_x - R_x) \quad (2)$$

where N_T is the total number of tuples in the database. For example N_b is obtained by $N_b = N_T - (S_b - R_b)$.

From Fig. 3, the $support$ and $confidence$ of rule $(A_1 = 1) \Rightarrow (A_2 = 1)$ is calculated as follows:

$$support((A_1 = 1) \rightarrow (A_2 = 1)) = R_b/N_b \quad (3)$$

$$confidence((A_1 = 1) \rightarrow (A_2 = 1)) = \frac{R_b/N_b}{R_a/N_a} \quad (4)$$

Important association rules are defined as the ones satisfying the following:

$$cosine > cosine_{min}, \quad (5)$$

$$support \geq sup_{min}, \quad (6)$$

$$confidence \geq conf_{min}, \quad (7)$$

$$confidence \geq support \quad (8)$$

$cosine_{min}$, sup_{min} and $conf_{min}$ are the minimum cosine, minimum support and minimum confidence values given by users. Table II shows an example of the measurements of some rules generated by node connections of Fig. 3.

Table II
EXAMPLE OF MEASUREMENTS OF ASSOCIATION RULES

Association Rule	*Support*	*Confidence*
$A_1 = 1 \rightarrow A_2 = 1$	$\frac{R_b}{N_b}$	$\frac{R_b/N_b}{R_a/N_a}$
$A_1 = 1 \rightarrow A_2 = 1 \wedge A_3 = 1$	$\frac{R_c}{N_c}$	$\frac{R_c/N_c}{R_a/N_a}$
$A_1 = 1 \rightarrow A_2 = 1 \wedge A_3 = 1 \wedge A_4 = 1$	$\frac{R_d}{N_d}$	$\frac{R_d/N_d}{R_a/N_a}$
$A_1 = 1 \wedge A_2 = 1 \rightarrow A_3 = 1$	$\frac{R_c}{N_c}$	$\frac{R_c/N_c}{R_b/N_b}$
$A_1 = 1 \wedge A_2 = 1 \rightarrow A_3 = 1 \wedge A_4 = 1$	$\frac{R_d}{N_d}$	$\frac{R_d/N_d}{R_b/N_b}$
$A_1 = 1 \wedge A_2 = 1 \wedge A_3 = 1 \rightarrow A_4 = 1$	$\frac{R_d}{N_d}$	$\frac{R_d/N_d}{R_c/N_c}$

The extracted important association rules are stored in a local pool all together through generations. When an important rule is extracted by GNP, the redundancy of the attributes is checked and it is also checked whether the important rule is new or not, that is, whether the rule is already in the local pool or not.

3) Genetic Operations: In order to extract important association rules it is necessary to change the connections of GNP-individuals. For instance, if the connection of P_1 is changed from node $A_1 = 1$ to node $A_3 = 1$ as shown in the lower part of Fig. 3, then, it is possible to calculate the support of $(A_3 = 1)$, $(A_3 = 1 \wedge A_4 = 1)$ and $(A_3 = 1 \wedge A_4 = 1 \wedge A_5 = 1)$ in the next examination.

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. These parameter values are chosen experimentally.

4) 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:

$$F = \sum_{r \in R} \{cosine(r) + \alpha_{new}(r) + \beta(NA_A(r) - 1) + \beta(NA_C(r) - 1)\} \quad (9)$$

The terms in Eq. (9) are as follows:

R: set of suffixes of extracted important association rules satisfying (5), (6), (7) and (8)

cosine(r): value of *cosine correlation measure* of rule r

$\alpha_{new}(r)$: additional constant defined by

$$\alpha_{new}(r) = \begin{cases} \alpha_{new} & \text{(rule } r \text{ is new)} \\ 0 & \text{(rule } r \text{ has been already extracted)} \end{cases} \quad (10)$$

β: coefficient for the number of attributes.

$NA_A(r)$: the number of attributes in the antecedent of rule r.

$NA_C(r)$: the number of attributes in the consequent of rule r.

Constants in Eq. 9 are defined empirically based on the values of $cosine(r)$. Thus, $\beta = 0.10$ and $\alpha_{new}(r) = 0.3$.

$NA_A(r) \leq MaxLength$ and $NA_C(r) \leq MaxLength$. $MaxLength = 2T + 1$, where T is the number of heterogeneous databases.

Cosine(r), $NA_A(r)$ and $NA_C(r)$, 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. Algorithm Summary

The algorithm for discovering general association rules from heterogeneous data with missing values can be summarized as follows:

INPUT: A dataset with n binary attribute values with missing values, a predefined number of generations T, a predefined minimum support (sup_{min}), minimum confidence ($conf_{min}$) and minimum cosine ($cosine_{min}$) thresholds.

OUTPUT: A pool of general association rules with support, confidence and cosine values larger than or equal to the predefined minimum $support$, $confidence$ and $cosine$ thresholds.

$STEP$ 1: Randomly generate a population of GNP individuals with a predefined number of judgment and processing nodes.

$STEP$ 2: Extract general association rules using GNP as follows:

$STEP$ 2.1: Evaluate if an attribute is missing or not using judgment nodes by the following: the transition from one judgment node to another is executed when the value is 1 or "m". Then go to the Continue-side of the judgment node, otherwise, go to the Skip-side of the judgment node.

$STEP$ 2.2: Calculate the rule measurements (support, confidence and cosine) using the number of available records on the Continue-side at each judgment using the processing nodes. That is, N_x, S_x and R_x.

$STEP$ 3: Check whether an important rule is new or not (whether it is already in the pool or not)

$STEP$ 4: Store the new general association rule that satisfy the minimum support, confidence and cosine thresholds.

$STEP$ 5: If the number of generations T reaches, then stop the algorithm, otherwise go to the next step.

$STEP$ 6: Perform the evolution of the GNP individuals as follows:

$STEP$ 6.1: Calculate the fitness of each GNP individual.

$STEP$ 6.2: Select the top $1/3$ GNP individuals according to their fitness values.

$STEP$ 6.3: Execute the genetic operators to the selected GNP individuals in order to create the next population.

$STEP$ 7: Go to $STEP$ 2.

V. Simulation Results

In order to test and validate the effectiveness of the proposed method, two real-time scientific databases from UCI ML Repository [17] and World Data System (WDS) [18] were taken to conduct the experiments, which are frequently used in data mining community. Both of them contains heterogeneous spatial-temporal data and they are suitable for mining general association rules. The first one ("A" dataset) is El Nino dataset and contains oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific. The second one ("B" dataset) correspond to the weather information of the Pacific Ocean taken by sensors of World Ocean Circulation Experiment (WOCE).

Table III shows the information of the original datasets, the first column of Table III shows the names of the datasets, the second shows the number of attributes, the third column shows the number of records, the fourth column shows if the dataset contains missing values and the fifth column shows the attribute characteristics of the dataset.

A. Experiment Setting

Both datasets are combined taken into account the date and each attribute is discretized into two corresponding attributes according to their values. For instance, if $Latitude \leq 0$ correspond to the $Latitude = South$. In this experiment, data only from one year (1993) is considered. After the discretization process, one large discretized dataset is generated, which contains 36 attributes and 20610 records. The combined dataset contains missing data, which varies for each attribute and ranges from 0% to 87%.

1) Parameters of GNP: The population size of GNP is 120. The number of processing nodes and judgment nodes in each GNP individual are 10 and 75, respectively. The maximum number of changing the connections of the processing nodes (*MaxLenght*) in each generation is $2(2) + 1 = 5$. The conditions of crossover and mutation are $P_c = 1/5$, $P_{m1} = 1/3$ and $P_{m2} = 1/5$. The termination condition T is 10, 30, 50 and 100 generations.

All algorithms were coded in Java language. Experiments were performed on a 3.2GHz Intel Xeon PC with 12G of main memory, running Windows 7 Ultimate 64bits.

Table IV shows some examples of the rules extracted by GNP. The termination "A" or "B" of each attribute means the correspondence to its dataset. From Table IV, the rules extracted by GNP are simple due to the small number of attributes in the antecedent part, which contribute to their understandability.

Fig. 5 shows the number of extracted rules when minimum confidence is 0.8 for different values of minimum support and number of generations. It can be seen that when the minimum support increases the number of rules extracted decreases for all generations because the constraints become more strict. Fig. 5 also shows that the number of rules increases when more generations in the evolution of GNP are used, especially at earlier generations.

Fig. 6 shows the number of extracted rules when the number of generations is 100 for different values of minimum support and minimum confidence. Fig. 6 shows that the minimum confidence has no great impact in the number of associations rules extracted compared with the minimum support.

 ISBN: 978-1-61208-162-5

Table III
INFORMATION OF THE ORIGINAL DATASETS

Dataset	*No. Attributes*	*No. Records*	Missing values	Attribute characteristics
El Nino	12	178080	Yes	Integer-Real
WOCE	14	71692	Yes	Integer-Real

Table IV
EXAMPLES OF RULES EXTRACTED BY GNP

Association Rules	Cosine
IF Air_Temp = High_A ∧ Longitude = East_B, THEN Longitude = West_A ∧ Speed = Low_B ∧ Temp_T_Air_C = Low_B	0.8327
IF Latitude=North_A ∧ Rel_Hum = High_B, THEN Longitude=West_A ∧ Speed=Low_B ∧ Pressure_Atm = Low_B ∧ Temp_T_Air_C = Low_B	0.8862
IF Sea_Surf_Temp=High_A ∧ Speed=Low_B ∧ Precip=High_B, THEN Pressure_Atm=Low_B ∧ Temp_Air=High_B	0.9179
IF Meridional_Winds=North_A ∧ Longitude=West_B ∧ Pressure_Atm=High_B, THEN Zon_Winds=West_A ∧ Rel_Hum=Low_B	0.9781
IF Latitude=South_A ∧ Temp_Water = High_B, THEN Longitude=West_A ∧ Zon_Winds = West_A ∧ Speed=High_B	0.8729
IF Zon_Winds = West_A ∧ Meridional_Winds = South_A ∧ Speed=Low_B, THEN Temp_T_Air_C = High_B	0.9297

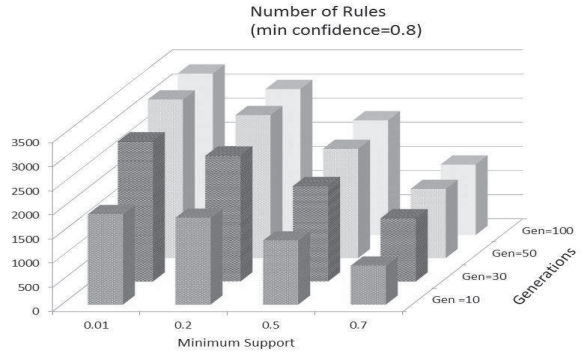


Figure 5. Number of extracted rules (min confidence=0.8)

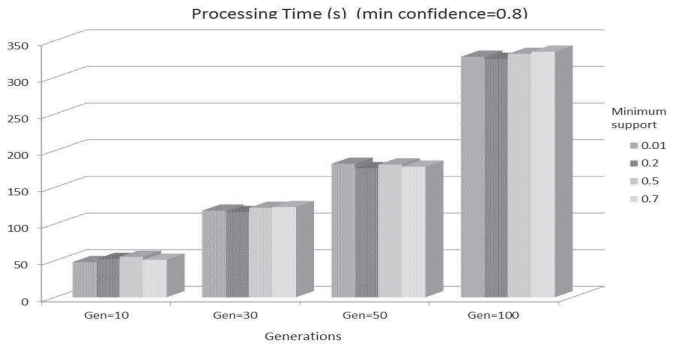


Figure 7. Processing Time (min confidence=0.8)

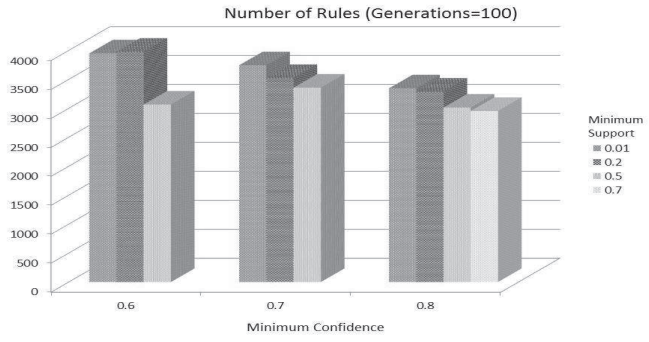


Figure 6. Number of extracted rules (generations=100)

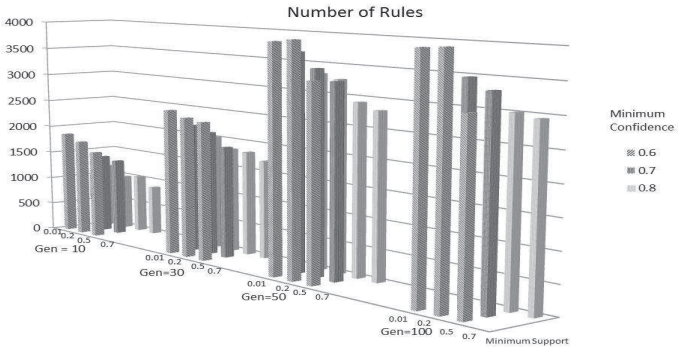


Figure 8. Number of Rules with different min support and min confidence

Fig. 7 shows the processing time for extraction of association rules when minimum confidence is 0.8 for different values of minimum support and number of generations. Fig. 7 shows that the processing time does not vary so much for a given generation as termination condition. On the other hand, the processing time increases when the number of generation increases because in every generation GNP searches and stores new association rules in the rule pool.

Fig. 8 shows the number of rules extracted with different conditions of minimum support, minimum confidence and the number of generations. Fig. 8 shows that although more

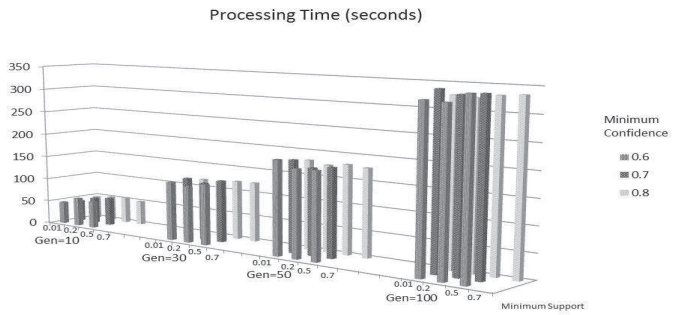


Figure 9. Processing Time with different min support and min confidence

association rules are extracted when used a larger number of generations, i.e., 100 generations; the difference, with the number of rules at 50 generations, is not so much. Therefore, most of the association rules are extracted in the earlier generations, which it is an advantage for the user's purpose.

The processing time increases when the number of generations are larger as shown in Fig. 9, however the number of extracted rules does not increase so much as it has been shown in Fig. 8. Therefore, 50 generations is enough in order to save processing time without risking of losing knowledge.

VI. Conclusion and Future Work

A method for association rule mining from incomplete databases has been proposed using GNP. An incomplete database includes missing data in some tuples, however, the proposed method can extract directly important rules using these tuples and users can define the conditions of important rules flexibly. The performance of the rule extraction has been evaluated using real data sets with a high rate of missing values. The results shows that the proposed method has the potential to realize associations considering heterogeneous databases and may be applied for rule discovery from incomplete databases in several other fields. For future work, the method may be extended to deal with large and heterogeneous databases with continuous values.

References

[1] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery, 8:53-87, 2004.

[2] A. K. H. Tung, H. Lu, J. Han, and L. Feng "Efficient Mining of inter-transaction association rules". IEEE Trans. on Knowledge and Data Engineering, 15(1): 43-56, 2003.

[3] A. Farhangfar, L. Kurgan, and J. Dy. "Impact of imputation of missing values on classification error for discrete data", *Journal of Pattern Recognition*, Vol 14, Issue 12, pp. 3692-3705, 2008.

[4] 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.

[5] 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.

[6] K. Shimada, K. Hirasawa, and J.Hu, "Genetic Network Programming with class association rule acquisition mechanism from incomplete databases". In *Proc. of the Society of Instrument and Control Engineers Annual Conference 2007*, pp. 2708-2714, 2007.

[7] K. Shimada and K. Hirasawa, "A method of Association Rule Analysis for Incomplete Database using Genetic Network Programming", in *Proc. of the Genetic and Evolutionary Computation Conference (GECCO 2010)*, pp.1115-5344, Portland, USA, 2010.

[8] C. Zhang and S. Zhang, *Association Rule Mining: models and algorithms*, Springer, 2002.

[9] C. C. Aggarwal and P.S. Yu . "A New Framework for Item Set Generation". *In: Proceedings of the ACM PODS Symposium on Principles of Database Systems*, pp. 18-24, Seattle, Washington (USA), 1998.

[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kauffman Publishers. USA, 2005.

[11] 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.

[12] 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.

[13] 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.

[14] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", in *Proc. of the 20th VLDB Conf.*, pp. 487-499, 1994.

[15] C.Z. Janikow, "A knowledge-intensive genetic algorithm for supervised learning", *Machine Learning 13*, pp. 189-228, 1993.

[16] C.C. Bojarczuk, H.S. Lopes, and A.A. Freitas, "Genetic programming for knowledge discovery in chest pain diagnosis", *IEEE Trans. on Engineering in Medicine and Biology Magazine*, Vol. 19, No.4, pp. 38-44, 2000.

[17] 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: Jun 14th, 2011]

[18] Walden, B; WOCE Surface Meteorology Data, WOCEMET (2006): Continuous meteorological surface measurement during KNORR cruise 316N138_12. Woods Hole Oceanographic Institution, Physical Oceanography Department.

 ISBN: 978-1-61208-162-5

New Solution for Extracting Inductive Learning Rules and their Post-Analysis

Rein Kuusik and Grete Lind
Department of Informatics
Tallinn University of Technology
Tallinn, Estonia
kuusik@cc.ttu.ee, grete@staff.ttu.ee

Abstract—**In this paper we present a new approach for machine learning task solution based on the new concept – the "Determinative set of rules" (DSR). We present a new inductive learning algorithm named MONSAMAX2 for finding DSR. MONSAMAX2 extracts it very effectively using some new pruning techniques. Compared to the former algorithm MONSIL it is much less labor-consuming. Also we present some ideas how to use DSR for post-analysis of rules.**

Keywords-machine learning; inductive learning algorithm; post-analysis of rules; determinative set of rules

I. INTRODUCTION

In the domain of inductive learning (IL) many different algorithms are used to solve different problems. There are several algorithms which try to solve the same task on different theoretical bases.

Some algorithms output rules as decision trees, some as set of rules, some of them find non-intersecting rules, some find overlapping rules, some algorithms find different systems of rules, some find a set of rules that meets certain requirements etc.

However, each method for finding a set of rules tries to prune the number of rules. This is expected, because the number of all possible rules in case of given sets of learning examples can be huge. Finding rules for bigger amounts of data is also very laborious. Because of this, we try to find such rules which have a stably good ability of recognition.

Thereby a number of different measures is used for evaluating the expediency of found rules. The main problem is not the existence of a rule for identifying some object but the correct identification of the object. Specifically as a result of using a rule set 3 situations can occur: 1) there is no rule for identifying the objects' belonging to a certain class, 2) the rule exists but identifies incorrectly, 3) the rule exists and it identifies correctly. The common approach uses the strategy that the first rule that identifies an object is used. But this does not guarantee certainty that the object is identified correctly. If several rules are used for identifying then the so-called rule conflict problem occurs. It occurs when two or more rules cover the same test example but predict different classes. The common strategy for preferring one of the conflicting rules is the best rule strategy [1]: for each rule a weight is calculated by some rule quality measure and the conflicting rule with the highest weight is chosen. Thereat the so-called preordering of rules [2] can be used for improving the result, as a consequence the result improves by 10-20%. It means that typically there are no actions for post-analysis of rules.

Also the approaches for the continuous development of the rule set have been created where the set of rules is complemented by adding new objects to the base of learning examples. Thereat the help of human experts can be used if the rule for identifying the object is missing [3].

It is clear that the pruning of the set of rules unavoidably leads to a loss of information and thereby to the possible increase of 1) the number of unidentified objects, 2) the number of misidentified objects. This paper offers one possible solution for lessening that.

The paper is structured as follows. The definitions of concepts used in inductive learning are given in Section II. Section III presents a proposed new approach for extracting rules. Conclusions are given in Section IV.

II. DEFINITIONS

We mainly follow the terms of the article [4].

The set of ***objects*** $X=\{x_1,\ldots,x_N\}$ can be described with ***attributes*** $t_1,\ldots, t_M$ so that every object x_i can be described as a tuple

$$x_i=(t_1(x_i),\ldots,t_M(x_i))=(x_{i1},\ldots,x_{iM}).$$

For each attribute t_j there exists a finite ***set of values*** A_j ($1\le j\le M$). So the ***attribute value*** x_{ij} of the object x_i belongs to the set A_j

$$x_{ij}=t_j(x_i)\in A_j.$$

Classes $C_1,\ldots,C_K$ are subsets of X such that

$$C_1\cup\ldots\cup C_K;\ \forall i\forall j,\ i\neq j : C_i\cap C_j=\varnothing.$$

The ***class value*** of the object $x\in X$ is c_j if $x_i\in Cj$. Let us denote the set of class values as

$$C=\{c_1,\ldots,c_K\}.$$

A ***learning example*** e_i is a tuple created from the object x_i and its class value

$$e_i=(x_i,c)=((t_1(x_i),\ldots,t_M(x_i)),c)=((x_{i1},\ldots x_{iM}),c).$$

Let us denote the ***set of examples*** E as

$$E=\{e_1,\ldots,e_n\}.$$

Let us denote the ***set of examples*** of class C_j as

$$E_j=\{e|e=(x,c_j),\ x\in C_j,\ C_j\subseteq X\}.$$

The ***positive example*** e_{j+} ***of the class*** C_j is an example which belongs to the set E_j, i.e.,

$$e_{j+}\in E_j\subseteq E.$$

The ***negative example*** e_{j-} ***of the class*** C_j is an example that does not belong to the set E_j, i.e.,

$$e_{j-}\notin E_j\subseteq E.$$

 ISBN: 978-1-61208-162-5

The ***majority class*** of set E is the class with the largest number of examples.

Function d which maps the class value c_j according to every element of the set X is called a ***concept***

$$d:X \rightarrow C;\ d(x_i)=c_j \Leftrightarrow x_i \in C_j.$$

In inductive learning the learning system should find, on the base of the learning examples, a concept description D which maps the class value for any object of the set X (not only for the objects of the example set) $D:X \rightarrow C$. Consequently the inductive learning system should (in an ideal case) find, on the base of learning examples, such a concept description D which maps the same class value as the concept d to every object

$$\forall x \in X,\ D(x)=d(x).$$

The ***(concept) description*** is the set of classification rules $D=\{r_1,\ldots,r_S\}$.

A ***classification rule*** (***decision rule***) is an implication where the condition part is a complex and the conclusion part is a class name:

$$r_j = \text{“}Com_j => c_k\text{”}$$

or

$$r_j = \text{“if } Com_j \text{ then } c_k\text{”}$$

or

$$r_j=(Com_j,c_k).$$

Complex Com_j is a tuple of ***selectors*** Sel_{jk} (k=1,…,M)

$$Com_j=(Sel_{j1},\ldots,Sel_{jM}).$$

Selector Sel_j is a subset of the set of values of the attribute t_j

$$Sel_j \subseteq A_j.$$

Description D ***maps a class value*** c_k ***for the object*** x_i if it contains a classification rule r_j which maps a class value c_k for the object x_i

$$\exists r_j \in D,\ r_j(x_i)=c_k \Rightarrow D(x_i)=c_k.$$

Rule $r_j=(Com_j,c_k)$ ***maps a class value*** c_k ***for the object*** x_i if its complex Com_j covers the object x_i

$$r_j=(Com_j,c_k),\ cover(Com_j,x_i) \Rightarrow r_j(x_i)=c_k.$$

Complex Com_j ***covers the object*** x_i if all its selectors Sel_{jk} cover this object

$$\forall k,\ 1 \le k \le M,\ cover(Sel_{jk},x_i) \Rightarrow cover(Com_j,x_i).$$

Selector Sel_{jk} ***covers the object*** x_i if the value of the attribute t_k of the object x_i is in the set Sel_{jk}

$$\forall j,\ 1 \le k \le M,\ x_{ik} \in Sel_{jk} \Rightarrow cover(Sel_{jk},x_i).$$

Description D is ***consistent*** on the set $X' \subseteq X$ if all its rules map the same class value for any object $x \in X'$

$$\forall r_i,r_j \in D,\ x \in X',\ X' \subseteq X,\ cover(Com_i,x),\ cover(Com_j,x) \Rightarrow r_i(x)=r_j(x).$$

Description D is ***complete*** on the set $X' \subseteq X$ if at least one rule for each object $x \in X'$ exists so that its complex covers this object

$$\forall x \in X',\ X' \subseteq X,\ \exists r_j \in D,\ cover(Com_j,x\}.$$

The inductive learning algorithms have to allow us to find descriptions that are at the same time both consistent and complete.

III. A NEW APPROACH

Next we present a new approach of IL which gives a new solution to previously named problems. At first we define a new concept “Determinative set of rules” (DSR), then describe an algorithm that can find it and describe how we can use this rule set for further analysis.

A. Basis of the New Approach

Let a data table X(N,M) be given and a set B of all possible rules for all classes and each rule in B is presented only once.

The Determinative set of rules (DSR) consists of all rules which are not contained in other rules of B.

B = {Ri}, i=1, 2, ..., K where K is a number of all possible rules. $Ri \neq Rj$, $i \neq j$.

$Ri \in DSR$ if there $/\exists\ Rt \in B, Ri \subset Rt, t \neq i$. $DSR \subseteq B$

It means that DSR does not contain the subrules of its rules. To get DSR from B we have to throw out all the subrules of the rules. We call this process „rule set compression“.

Example. Let B contain 4 rules:

r1: IF T1=1 & T2=1 THEN CLASS=1

r2: IF T1=1 & T3=2 THEN CLASS=2

r3: IF T2=1 THEN CLASS=1

r4: IF T3=2 THEN CLASS=2

As we see, the rule r1 is contained in r3 and r2 is contained in r4. According to the definition $DSR_B = \{r3, r4\}$.

The main features of DSR are:

1. there are no redundant attributes in rules,
2. the same object in X can be described by several rules.

B. Description of the Algorithm

Here we describe the algorithm realizing the new IL approach. The findable set of rules is DSR together with some redundant rules which are eliminated afterwards (rule set compression).

Algorithm MONSAMAX2 is given in Fig. 1.

This is a depth-first-search algorithm that makes subsequent extracts of objects containing certain factors (i.e., an attribute with a certain value). At each level first the rules (of that extract) are detected and then factors for making extracts of the next level are selected one by one.

The algorithm uses frequency tables for both all the objects in the current extract and each class of the current extract. We call them “3D frequency tables”. If a factor has equal frequencies for all objects and in any of the classes then this factor completes a rule. The rule includes also the factors chosen on the way to that extract.

The selection criteria for choosing the next factor are based on frequencies, the maximal frequency for all objects (of extract). If only one attribute (of the extract) has free (unused) value(s) (indicated by frequencies over zero) then it is not practical to make a next (further) extract because there would be no free factors to distinguish objects of different classes in that extract. If there are no free factors (i.e., no frequencies over zero) then obviously it is not possible to make a next extract. In both cases the algorithm backtracks to the previous level.

 ISBN: 978-1-61208-162-5

Algorithm MONSAMAX2
S0. t:=0; U_t:=∅
S1. Find frequencies in whole dataset and each class
If t>0 then
Bring zeroes down
Backward comparison
S2. For each factor A such that its frequency in some class C is equal to its frequency in the whole set
output rule {Ui}&A→C, i=0,…,t
A←0
S3. If not enough free factors for making an extract then
If t=0 then Goto End
Else t:=t-1; Goto S3
S4. Choose a new (free) factor U_t
U_t ←0; t:=t+1;
extract subtable of objects containing U_t;
Goto S1
End. Rules are found

Figure 1. Algorithm MONSAMAX2

Each factor that has been used for making an extract or completing a rule is set to zero in the corresponding frequency table. Zero in the frequency table means that this factor is eliminated from the analysis at this level.

Each frequency table (except for the initial level) inherits all zeroes of the previous level (we call it "bringing zeroes down").

Also, after making an extract, its (non-zero) frequencies are compared to the ones of the previous level. Equal frequencies at both levels mean that all objects containing that factor are contained in the extract of the current level and all possible rules containing them are found at current and subsequent levels. In order to prevent repetitious finding of such rules the frequencies of those factors are set to zero at the previous level. This technique is called "backward comparison". Using this pruning technique we can also determine the extractedness of all rules for some class, i.e., if for some class all frequencies are equal to zero at the initial level, it means that all rules for this class are found.

All these techniques can effectively decrease the number of extracts (nodes of the search tree) without losing the rules of DSR.

C. Example

In the following example data from [5] are used (Table I). In order to get a numerical representation the coding shown in Table II is used. Coded data are shown in Table III.

For given data frequencies are found across all data and across each class (see Table III). If frequencies of some factor are equal in the whole dataset and some class, we can complete the rule. In the given dataset/extract that factor determines the class. From the initial frequency tables (Table III) 3 rules are found this way:

R1: T2.1 → Class 1
R2: T3.2 → Class 1
R3: T2.2 → Class 2

TABLE I. EXAMPLE SET (FROM QUINLAN)

Object	Height	Hair	Eyes	Class
1	tall	dark	blue	–
2	short	dark	blue	–
3	tall	blond	blue	+
4	tall	red	blue	+
5	tall	blond	brown	–
6	short	blond	blue	+
7	short	blond	brown	–
8	tall	dark	brown	–

TABLE II. CODING OF VALUES

Attribute Value	Height T1	Hair T2	Eyes T3	Class
1	short	dark	blue	–
2	tall	red	brown	+
3		blond		

TABLE III. INITIAL DATA AND FREQUENCIES

Object	T1	T2	T3	Class
1	*2*	*1*	*1*	*1*
2	*1*	*1*	*1*	*1*
3	*2*	*3*	*1*	*2*
4	*2*	*2*	*1*	*2*
5	*2*	*3*	*2*	*1*
6	*1*	*3*	*1*	*2*
7	*1*	*3*	*2*	*1*
8	*2*	*1*	*2*	*1*

Value	T1	T2	T3	Class
1	3	**3**	5	
2	5	**1**	**3**	*all*
3		4		
1	2	**3**	2	
2	3	0	**3**	*1*
3		2		
1	1	0	3	
2	2	**1**	0	*2*
3		2		

The frequencies of those factors (T2.1, T2.2, T3.1) are set to zero in the current frequency table (see Table IV). Now the factor with the biggest frequency is selected for making an extract. We have two candidates: T1.2 and T3.1, both with frequency 5. As we do not have additional information we choose the first one. The chosen factor is T1.2 (with frequency 5). The extract by T1.2 and the corresponding frequencies are given in Table V.

The cells with grey backgrounds are prohibited factors that have zeroed frequencies in the previous level. This frequency table completes no rules. T3.1 with frequency 3 is chosen for making a subsequent extract (see Table VI).

In this frequency table the frequency of T2.3 in Class 2 is the same as in the previous level (see Table V), in the

previous level this frequency is set to zero (because everything connected to it will be done at a lower level).

From the current frequency table the next rule is found:

R4: T1.2&T3.1&T2.3 → Class 2

After completing a rule, the frequency of T2.3 is set to zero and the current frequency table contains no more usable frequencies.

Turning back to the previous level (Table V) it occurs that after zeroing the frequency of T3.1 (as a basis of the extract just made) there is only one usable factor (T2.3). It makes no sense to make an extract by it.

Therefore we turn back to the initial level. The frequencies are given in Table VII. The frequency of the last basis for the extract T1.2 is set to zero. The basis for the next extract is T3.1 with frequency 5. The extracted data and corresponding frequencies are given in Table VIII.

Backward comparison finds two factors with equal frequencies at the current and previous (see Table VII) levels: T1.1=1 in Class 2 and T2.3=2 in Class 2. Both frequencies are set to zero at the previous level. As we can see, the frequency table for Class 2 at the initial level is empty which means that all the rules for Class 2 will be extracted after traversing the extract by T3.1.

TABLE IV. FREQUENCIES AFTER EXTRACTING 3 RULES

Value	T1	T2	T3	Class
1	3	*0*	5	
2	**5**	*0*	*0*	*All*
3		4		
1	2	*0*	2	
2	**3**	*0*	*0*	*1*
3		2		
1	1	*0*	3	
2	**2**	*0*	*0*	*2*
3		2		

TABLE V. EXTRACT BY T1.2=5 AND CORRESPONDING FREQUENCIES

Object	T1	T2	T3	Class
1		*1*	*1*	*1*
3		*3*	*1*	*2*
4		*2*	*1*	*2*
5		*3*	*2*	*1*
8		*1*	*2*	*1*

Value	T1	T2	T3	Class
1		*0*	**3**	
2		*0*	*0*	*all*
3		2		
1		*0*	**1**	
2		*0*	*0*	*1*
3		1		
1		*0*	**2**	
2		*0*	*0*	*2*
3		1		

TABLE VI. EXTRACT BY T1.2&T3.1=3 AND CORRESPONDING FREQUENCIES

Object	T1	T2	T3	Class
1		*1*		*1*
3		*3*		*2*
4		*2*		*2*

Value	T1	T2	T3	Class
1		*0*		
2		*0*		*All*
3		**1**		
1		*0*		
2		*0*		*1*
3		0		
1		*0*		
2		*0*		*2*
3		**1**		

TABLE VII. FREQUENCIES AT THE INITIAL LEVEL

Value	T1	T2	T3	Class
1	3	*0*	**5**	
2	*0*	*0*	*0*	*All*
3		4		
1	2	*0*	**2**	
2	*0*	*0*	*0*	*1*
3		2		
1	**1**	*0*	**3**	
2	*0*	*0*	*0*	*2*
3		**2**		

TABLE VIII. EXTRACT BY T3.1=5 AND CORRESPONDING FREQUENCIES

Object	T1	T2	T3	Class
1	*2*	*1*		*1*
2	*1*	*1*		*1*
3	*2*	*3*		*2*
4	*2*	*2*		*2*
6	*1*	*3*		*2*

Value	T1	T2	T3	Class
1	2	*0*		
2	*0*	*0*		*all*
3		**2**		
1	1	*0*		
2	*0*	*0*		*1*
3		0		
1	**1**	*0*		
2	*0*	*0*		*2*
3		**2**		

From the current extract (Table VIII) we get a rule:

R5: T3.1&T2.3 → Class 2

After the frequency of T2.3 is set to zero (at the current level) only one non-zero frequency is left (for T1.1). The

 ISBN: 978-1-61208-162-5

possible extract by it cannot give any rules. Therefore algoritm backtracks to the initial level.

The current state of frequencies is given in Table IX.

The next extract is made by T2.3 (see Table X), there are no rules. There is only one frequency above zero in the frequency table, therefore we backtrack to the previous (initial) level.

In the frequency table of the initial level (see Table XI) there is now only one usable (non-zero) frequency. It cannot give a rule because if it could then it could be extracted from the initial table at the beginning. An extract is not made. The work is finished.

TABLE IX. FREQUENCIES AT THE INITIAL LEVEL

Value	T1	T2	T3	Class
1	3	*0*	*0*	
2	*0*	*0*	*0*	*all*
3		**4**		
1	2	*0*	*0*	
2	*0*	*0*	*0*	*1*
3		**2**		
1	*0*	*0*	*0*	
2	*0*	*0*	*0*	*2*
3		***0***		

TABLE X. EXTRACT BY T2.3=4 AND CORRESPONDING FREQUENCIES

Object	T1	T2	T3	Class
3	*2*		*1*	*2*
5	*2*		*2*	*1*
6	*1*		*1*	*2*
7	*1*		*2*	*1*

Value	T1	T2	T3	Class
1	2		*0*	
2	*0*		*0*	*all*
3				
1	1		*0*	
2	*0*		*0*	*1*
3				
1	*0*		*0*	
2	*0*		*0*	*2*
3				

TABLE XI. FREQUENCIES AT THE INITIAL LEVEL

Value	T1	T2	T3	Class
1	**3**	*0*	*0*	
2	*0*	*0*	*0*	*all*
3		*0*		
1	**2**	*0*	*0*	
2	*0*	*0*	*0*	*1*
3		*0*		
1	***0***	*0*	*0*	
2	*0*	*0*	*0*	*2*
3		*0*		

So, we extracted 5 rules: R1: T2.1 → Class 1, R2: T3.2 → Class 1, R3: T2.2 → Class 2, R4: T1.2&T3.1&T2.3 → Class 2, R5: T3.1&T2.3 → Class 2.

As we see the extracted rule set is not DSR because of the rule R4 which is a subrule of R5. After the compression of the extracted rule set we get a DSR: R1, R2, R3 and R5.

The number of extracted rules for MONSAMAX2 depends on the criteria of choosing the leader value for an extract in a situation when there are several candidates with equal frequencies. For example, if we would choose in the beginning of the algorithm T3.1 (with frequency 5) as a leader value instead of T1.2 (frequency=5), then we would extract only 4 rules (R1, R2, R3 and R5) and compression would not be needed (but we do not know this). MONSAMAX2 produces more additional information for effective rule set compression, but here is not enough space for presenting it.

D. Discussion

For the same purpose – to find a complete and consistent description – algorithms MONSIL [6] and DEILA [7] have been proposed. Each of the algorithms (MONSAMAX2, MONSIL and DEILA) work in a different way and usually give different descriptions. The common idea is the step following the main algorithm – compression of the found rule set in order to get a result as compact as possible .

Similarly to MONSAMAX2 the result of MONSIL (before compression) depends on the choice of leader value for making extract when there are several candidates with equal frequencies. For the same Quinlan's data [5] as here (see Table I), two different results of MONSIL are given (in [6]): the first one consisting of 8 rules and the second – 5 rules. In the latter, the redundant rule (T1.1&T2.3&T3.1 → Class 2) is not the same as in case of MONSAMAX2 (T1.2&T3.1&T2.3 → Class 2). The result which consists of 8 rules contains two more redundant rules (containing T1.1) in addition to these two.

We noticed that after compression the results of MONSIL and MONSAMAX2 are the same. This rule set is called DSR (determinative set of rules).

Algorithm DEILA finds more rules than MONSAMAX2 and MONSIL. From Quinlan's data it finds 15 rules. The result of DEILA may not contain all DSR rules. Some of them can be "replaced" by longer rules. This is due to DEILA's working principle – all found rules are dicliques.

The amount of extracted rules for MONSAMAX2 is smaller than for MONSIL because the first one extracts shorter rules first while the latter extracts longer rules first. The difference is in the number of (redundant) subrules – the rules that will be removed by compression. MONSAMAX2 finds fewer such rules (due to finding shorter rules first).

In order to determine the belonging of the extracted objects to the same class in the process of extracting rules, MONSIL must make an extract and usually the objects do not belong to the same class. It means that we have made a superfluous effort. MONSAMAX2 works so effectively because we have data to determine belonging of objects to the same class using 3D frequency tables, there is no need to make these extracts.

 ISBN: 978-1-61208-162-5

During the work of MONSAMAX2 we can also observe every class covering with rules: if 3D frequencies for some class are empty at the initial level it means that all rules for this class are extracted.

IV. CONCLUSION

This paper proposes a new approach and the corresponding algorithm MONSAMAX2 for finding a determinative set of overlapping rules. The algorithm is based on frequency tables and new pruning techniques which make it easy to detect a potential DSR rule.

MONSAMAX2 is more effective compared to the former algorithm MONSIL because it prevents the making of many unnecessary extracts due to using 3D frequency tables. Also it finds less redundant (i.e., non-DSR) rules because it finds shorter rules first while MONSIL starts from the longer ones.

On the basis of DSR we can form and solve next tasks, for example, to find

1. the shortest rules (by the number of attributes (selectors) in the rule),
2. the longest rules (by the number of attributes in the rule),
3. the rules with specific features (for example, all rules with r selectors),
4. the shortest rule system (i.e., the rule system with the smallest number of rules),
5. the rule system which consists of rules with minimal number of selectors,
6. all the rule systems we can form on the basis of DSR.

All these tasks are necessary for the post-analysis of the extracted rules. It means that several new possibilities are available for experimentation with several rule sets (subsets of DSR) and for describing them. We must not try to minimize the rule set during the work of a machine learning algorithm, we can find the best solution during the post-analysis of DSR.

Using DSR and the post-analysis of rules also gives the possibility to gather statistics about the use of rules in classification in order to analyze the rules' perspective and their power of classification. We can also see which rules classify more accurately and which do not on the basis of the information we have about classified (test-set and real) objects. On this basis we can reorder the rules in the rule set. DSR is a good basis for developing this approach.

Somebody might say that the finding of DSR is very laborious, especially in cases of large amounts of data. If so, the user can decide what is the purpose of the work. If the purpose is a quick one-time information gathering for a data set under analysis then the use of DSR-based IL approach may not be the best one. But if the purpose is to describe the data set and through that discover new knowledge and get an opportunity for post-analysis of the rule set then this approach is a good solution.

The post-analysis of rules will be the topic of the next paper.

REFERENCES

[1] L. Torgo, "Rule combination in inductive learning," in Machine Learning: ECML-93, ser. Lecture Notes in Computer Science, P. Brazdil, Ed. Springer Berlin / Heidelberg, 1993, vol. 667, pp. 384-389.

[2] T. Treier, "A new effective approach for solving the rules conflict problem", 2011 International Conference on Intelligent Computing and Control (ICOICC 2011), May 2011, in press.

[3] I. Birzniece, "From Inductive Learning towards Interactive Inductive Learning," in Scientific Journal of Riga Technical University, Computer Science. Applied Computer Systems, vol. 41, 2010, pp. 106-112.

[4] M. Gams and N. Lavrac, "Review of Five Empirical Learning Systems within a Proposed Schemata," in I. Bratko, N. Lavrac (Eds.), Progress in Machine Learning, Proceedings of EWSL 87: 2nd European Working Session on Learning, Bled, Yugoslavia, May 1987. Sigma Press, Wilmslow, 1987, pp. 46-66.

[5] J. R. Quinlan, "Learning efficient classification procedures and their application to chess end games," in J. G. Carbonell, R. S. Michalski, T. M. Mitchell (Eds.), Machine Learning. An Artificial Intelligence Approach, Springer-Verlag, 1984, pp. 463-482.

[6] P. Roosmann, L. Võhandu, R. Kuusik, T. Treier, and G. Lind, "Monotone Systems approach in Inductive Learning," in International Journal of Applied Mathematics and Informatics, Issue 2, Vol. 2, 2008, pp. 47-56.

[7] R. Kuusik, T. Treier, G. Lind, and P. Roosmann, "Machine Learning Task as a Diclique Extracting Task," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery: FSKD'09, Tianjin, China, August 14-16, 2009; Los Alamitos, California: Conference Publishing Service, 2009, pp. 555-560.

 ISBN: 978-1-61208-162-5

An Equivalence Class Based Clustering Algorithm for Categorical Data

Liu Qingbao, Wang Wanjun, Deng Su
Sci. and Technol. on Inf. Syst. Eng. Laboratory
National University of Defense Technology
Changsha, China, 410073
e-mail: liuqingbao@nudt.edu.cn,
wwjunbelieve@gmail.com, Sudeng@sohu.com

Guozhu Dong
Department of Computer Science & Engineering
Wright State University
Dayton, Ohio, USA 45435
e-mail: guozhu.dong@wright.edu

***Abstract*—Categorical data clustering plays an important role in data mining for many applications, including popular applications involving text mining and blog mining. While most traditional clustering methods rely on a distance function. However, the distance between categorical data is hard to define, especially for exploratory situations where the data is not well understood. As a result, many clustering methods do not perform well on categorical datasets. In this paper we propose a novel Equivalence Class based Clustering Algorithm for Categorical data (ECCC). ECCC takes the support transaction sets of selected frequent closed patterns as the candidate clusters. We define a novel quality measure to evaluate the suitability of frequent closed patterns to form the clusters; the measure is based on two factors: cluster coherence expressed in terms of closed patterns, and cluster discriminativeness expressed in terms of quality and diversity of minimal generator patterns. ECCC uses that measure to select the high quality frequent closed patterns to form the final clusters.**

Keywords-clustering analysis; categorical data; equivalence class

I. INTRODUCTION

Clustering is unsupervised and highly explorative. It is an important approach, widely used in life science, medicine, social science, engineering and many other fields [1]. Most traditional clustering methods rely on a distance function. Since the distance between categorical data is hard to define, especially for exploratory situations where the data is not well understood, many clustering methods do not work well on categorical datasets.

Several clustering methods have been recently developed to handle categorical data, including k-ANMI introduced in [2], Squeezer proposed in [3], GAClust in [4], ccdByEnsemble in [5], the Entropy-based algorithm in [6], and ECCLAT in [7]. However, these methods have various shortcomings; for example, the Entropy-based algorithm emphasizes intra-cluster purity while ignoring inter-cluster separation.

Here we detail the ECCLAT algorithm, which we will compare our method with. Firstly, it is necessary to explain the term "frequent closed itemset" in a general manner: a closed itemset is a maximal set of items shared by a set of transactions; when the frequency of a closed itemset is larger than the frequency threshold (denoted as: *minfr*), it is called a frequent closed itemset. ECCLAT extracts a subset of concepts from the lattice of frequent closed itemsets, using the evaluation measure *interestingness(X) = (homogeneity(X) + concentration(X))/2* [7] (*X* is an itemset). More specifically, ECCLAT first mines the set of frequent closed itemsets; it views the support transaction set of each frequent closed itemset as a candidate cluster. Then, it computes the interestingness of the candidate clusters, and iteratively selects the next candidate cluster having the highest interestingness as the next final cluster. ECCLAT is an approximate clustering algorithm and allows the clusters to have transaction overlap. The *concentration(X)* measure is defined to limit the overlap of transactions between clusters by taking into account the number of candidate clusters where each transaction of *X* appears. Checking the definition of *concentration(X)* in [7], we can see that it treats all candidate clusters as equally important. Such weaknesses of ECCLAT lead to poor clustering results.

In this paper, we propose a novel method called ECCC (an Equivalence Class based Clustering Algorithm for Categorical data). The main contribution of the paper is to provide a better quality measure to replace the concentration quality measure of ECCLAT. Our quality measure, called inter-cluster discriminativeness index, will consider the quality and diversity/richness of the minimal generator patterns. Specifically, our algorithm first mines the equivalence classes [9] of patterns, including the closed patterns and their associated sets of generator patterns. Similarly to ECCLAT, we also regard the support transaction set of each closed pattern as a candidate cluster. We combine the intra-cluster homogeneity index of ECCLAT and our inter-cluster discriminativeness index into a general and objective quality index on clusters. Our algorithm then selects the high quality clusters from the candidate clusters using that quality index.

Compared against ECCLAT, our ECCC uses both the closed pattern and the generator patterns, instead of just the closed patterns, of equivalence class to define the *discriminativeness index.* ECCC prefers the equivalence classes that have a long closed pattern and many short generator patterns. The first advantage of ECCC is that it avoids the drawback of ECCLAT mentioned above. The second advantage is that there is no transaction overlap among the final clusters. The third advantage is that ECCC needs only one parameter (*minfr)* while ECCLAT needs two parameters (*minfr* and *M* [7]*).* As a result, ECCC is more accurate in recovering expert defined classes than ECCLAT.

In Section II, we describe our discriminativeness index and ECCC algorithm, after giving relevant preliminaries. In Section III, we report experimental results. Our conclusions are presented in Section IV.

Supported by the National Natural Science Fund of China under Grant No.70771110.

 ISBN: 978-1-61208-162-5

II. ECCC

In this section we will present our method, namely ECCC, after firstly giving some definitions.

A. Preliminaries

We assume that we are given a dataset D (a set of transactions) in the following definitions. For each itemset z, let $f_D(z)=\{t \in D | z \subseteq t\}$ denote the support set of transactions for z.

Definition 1 (*EC: Equivalence Class***)** [9]. An equivalence class *EC* is a (maximal) set of frequent itemsets (also called frequent patterns) that have a common support set of transactions.

So, if *EC* is an equivalence class and x and $y \in EC$, then $f_D(x)=f_D(y)$.

Here we give the definition of "frequent closed itemset" which is called "closed pattern" in our ECCC.

Definition 2 (*cp: closed pattern***)** [9]. Given an equivalence class *EC*, the closed pattern *cp* of *EC* is

$$cp = \bigcup_{p \in EC} p \,. \tag{1}$$

Definition 3 (*gp: generator pattern***)** [9]. Given an equivalence class *EC*, a pattern $gp \in EC$ is a generator pattern of *EC* if, for $\forall\ z \in EC$ *s.t.* $z \neq gp$, it is the case that $z \not\subset gp$.

It is well known that an equivalence class has only one closed pattern and it has one or more generator patterns. We will represent an equivalence class *EC* as *EC*=[*G(cp),cp*] [9], where $G(cp)=\{gp_i | 1 \le i \le k\}$ which is the set of generator patterns and cp is the closed pattern of EC.

Definition 4 (*candidate cluster***)**. A set of transactions $CC \subseteq D$ satisfying $CC=f_D(cp)$ for some closed pattern cp is called a candidate cluster associated with *cp*; this CC will be denoted by *CC(cp)*.

B. Homogenity and Discriminativeness Measures

There are often many candidate clusters, and only a few candidate clusters can become the final clusters. For example, with *minfr* = 5%, there are 9738 candidate clusters in the mushroom dataset. So we need quality measures to select the candidate clusters as final clusters. Our algorithm will use one factor's formula used by ECCLAT, and replaces the other factor's formula using a new one.

Definition 5 (*HI: Homogeneity Index***)**. The Homogeneity Index [7] of a candidate cluster *CC(cp)* is defined by:

$$HI_{CC}(cp) = \frac{|CC(cp)| \times |cp|}{divergence(cp) + |CC(cp)| \times |cp|} \tag{2}$$

where $divergence(cp) = \sum_{t \in f_D(cp)} |t - cp|$, and |S| denotes the cardinality of a set S.

Homogeneity Index is used to measure the intra-cluster similarity. Larger values are better. Using this index, we prefer those candidate clusters whose closed patterns are very long. If a candidate cluster *CC(cp)* has a very long closed pattern *cp*, then all the transactions in *CC(cp)* share all items in *cp*, implying that *CC(cp)* is highly coherent; we note that *divergence(cp)* is small and $HI_{CC}(cp)$ is large in this situation.

For the inter-cluster diversity, we propose a novel measure called *Discriminativeness Index* which is defined below.

Definition 6 (*DI: Discriminativeness Index***)**. The *discriminativeness index* of a candidate cluster *CC(cp)* is defined as:

$$DI_{CC}(cp) = \prod_{gp_i \in G(cp)} (1 + \frac{|cp - gp_i|}{|cp|}) \tag{3}$$

where $|cp|$ and $|cp\text{-}gp_i|$ are the number of items in *cp* and $cp\text{-}gp_i$, respectively.

Larger $DI_{CC}(cp)$ values are better. Using this *Discriminativeness Index*, we prefer the candidate cluster which has a very long closed pattern and many short generator patterns. Our rationale for *Discriminativeness Index* is similar to that in [10]. If a candidate cluster *CC(cp)* has *many short* generator patterns, then each such short generator pattern *gp* is a strong discriminator that can be used to easily separate and distinguish *CC(cp)* from the other candidate clusters. The shorter gp is the easier it is to do the separation. The more such short gp patterns we have, the more different ways we have to describe the cluster and discriminate it from other clusters. So we think the high $DI_{CC}(cp)$ value implies that this candidate cluster *CC(cp)* is significantly different from other candidate clusters, is identified very easily, and has better quality.

Definition 7 (*QI: Quality Index***)**. The EC based Quality Index of the candidate cluster *CC(cp)* is defined as follows:

$$QI_{CC}(cp) = HI_{CC}(cp) \times DI_{CC}(cp) \,. \tag{4}$$

Our idea is to select these candidate clusters with high *Quality Index* as the final clusters. The next section presents an algorithm for this task.

C. The Process of ECCC

On dataset *D*, we first use DPMiner algorithm [9] to mine the closed patterns and their generators simultaneously, using a minimal frequency threshold *minfr*. Then, we determine the candidate clusters of the frequent closed patterns, calculate the quality of each candidate cluster, and select the candidate cluster *CC(cp*)* with highest quality as a final cluster *C(cp*)*. When there are two and more highest quality candidate clusters, we prefer the candidate cluster with larger number of transactions. For any remaining candidate cluster *CC(cp)* such that $cp \neq cp^*$ and $CC(cp) \bigcap C(cp^*) \neq \varnothing$, we modify the candidate cluster *CC(cp)* as *CC(cp)=CC(cp)-C(cp*)*. If $|CC(cp)| < minfr$, we delete the candidate cluster *CC(cp)*. Then we recalculate $HI_{CC}(cp)$, $DI_{CC}(cp)$ and $QI_{CC}(cp)$ of the

 ISBN: 978-1-61208-162-5

candidate clusters, and select the candidate cluster *CC(cp*)* with highest quality as the next final cluster. We repeat the process above, until there is no candidate cluster. At the end, we classify all remaining transactions into the trash set.

The *pseudo-codes* of ECCC are given below.

Input:
D is a dataset to be clustered;
minfr is the frequency threshold;
Output:
CL is the set of Clusters;
Trash is the set of the trash transactions;
Description:

1. mine $CP=\{cp_k|1\le k\le N\}$, $G(cp_k)$, $CC(cp_k)$;
2. **for** each $cp\in CP$ **do**
3. calculate $HI_{CC}(cp)$ *and* $DI_{CC}(cp)$;
4. $QI_{CC}(cp)= HI_{CC}(cp)\times DI_{CC}(cp)$;
5. **end for**
6. select $CC(cp^*)$, *s.t.* $QI_{CC}(cp^*)= \max_{cp_k\in CP}\{QI_{CC}(cp_k)\}$;
7. $C(cp^*)= CC(cp^*)$;
8. delete $CC(cp^*)$;
9. insert $C(cp^*)$ into *CL*;
10. **for** each $CC(cp) \wedge (CC(cp)\cap C(cp^*)\neq \varnothing)$ **do**
11. $CC(cp)=CC(cp)-C(cp^*)$;
12. **if** $|CC(cp)|<minfr$ **then**
13. delete $CC(cp)$;
14. **else**
15. recalculate $HI_{CC}(cp)$, $DI_{CC}(cp)$ and $QI_{CC}(cp)$;
16. **end if**
17. **end for**
18. repeat *steps 6--17* until there is no candidate cluster;
19. classify the remaining transactions of *D* into the *Trash*;
20. return *CL* and *Trash*;

III. EXPERIMENT RESULTS

We now use experiments to demonstrate that (1) the ECCC algorithm is accurate and (2) the ECCC algorithm is scalable.

Experiments were conducted on a desktop computer with a 2.33 GHz Intel CPU and 3 GB memory running the Windows XP.

A. Accuracy test on Mushroom Dataset and Zoo Dataset

We evaluate the accuracy of our algorithm ECCC on two real datasets available at the UCI Repository [8]. One is the well known mushroom dataset which has 22 attributes, 8124 transactions, and two class labels provided by domain experts. The other is the zoo database (101 transactions with 7 class labels provided by domain experts) which has 15 boolean attributes and a numerical one ("legs"), we used the six values of "legs" as categorical values.

1) Error rates test on mushroom dataset

In this section, we present experiment results of ECCC on the mushroom dataset.

a) Compare against ECCLAT algorithm

For *minfr* = 5% and *M* = *minfr*, ECCLAT obtains 16 clusters and a trash cluster with slight overlapping between clusters 14 and 16 [7]. Also ECCC can obtain 16 clusters and a trash cluster without overlapping when *minfr* = 4%.

The comparison between the above two clusterings is shown in Table I. It is obvious that the ECCC clustering errors are lower than that of the ECCLAT clustering.

TABLE I. COMPARISON BETWEEN ECCC AND ECCLAT

Cluster No.	ECCC (*minfr* = 4%)		ECCLAT (M=*minfr* = 5%)	
	#Poisonous	***#Edible***	***# Poisonous***	***# Edible***
1	0	576	0	432
2	432	0	0	432
3	0	384	0	432
4	0	384	0	432
5	864	0	648	0
6	576	0	648	0
7	576	0	432	0
8	576	0	432	0
9	0	400	432	0
10	0	400	432	0
11	272	***96***	0	768
12	***72***	528	0	512
13	***128***	384	352	***96***
14	0	384	***288***	896
15	***144***	384	0	416
16	240	***96***	***72***	560
Trash	***36***	192	180	***160***
Error	***572***		***616***	

b) Average clustering error rates comparison

To further test the clustering errors of ECCC, we repeated ECCC with 10 different *minfr*s from 1% to 10% on mushroom dataset. The 10 results are shown in Table II.

TABLE II. RESULTS IN TERM OF THE DIFFERENT *MINFRS*

***minfr*(%)**	**#Clusters (including the trash cluster)**	**# Errors**
1	28+1	252
2	24+1	252
3	20+1	172
4	16+1	572
5	12+1	890
6	11+1	890
7	11+1	890
8	6+1	890
9	6+1	890
10	5+1	890

 ISBN: 978-1-61208-162-5

Table II shows that the clustering errors change with different *minfr*s. For *minfr* =3%, the result is the best. And for *minfr* =4%, the result is the middle. But when *minfr* changes from 5% to 10%, the errors do not change. So we can think the ECCC is a stable clustering algorithm.

In addition, we used the EM algorithm [11] which is implemented in WEKA [12] to generate clustering for comparison against the ECCC. Results are given in Table III.

TABLE III. AVERAGE CLUSTERING ERROR RATES COMPARISON

Algorithm	Average Clustering Error Rates
ECCC	0.081
EM	0.133
k-ANMI	0.165
ccdByEnsemble	0.315
GAClust	0.393
Squeezer	0.206

Table III indicates that the average clustering error rate in Table II is lower than that of EM and the algorithm of [2].

2) Purity test on zoo dataset

We now report experiments on the zoo database, to demonstrate that ECCC is accurate with high purity, and to compare it against Entropy-based algorithm and K-means algorithm. Table IV indicates that our ECCC is better than Entropy-based algorithm and K-means algorithm on purity. (The purity of a clustering (C_1,…,C_m) against an expert given clustering (C'_1,…,C'_k) is defined as follows: For each cluster C_i, let C'_{i*} denote the expert cluster with the largest overlap with C_i. Purity of C_i is defined as $|C_i \bigcap C'_{i*}| / |C_i|$. The purity of the clustering (C_1,…,C_m) is defined as the weighted average of the purity of the clusters.)

TABLE IV. RESULTS COMPARISON ON ZOO DATASET

	ECCC	Entropy-based	K-means
Purity	0.9208	0.9000	0.8400

In summary, these experimental results on both mushroom dataset and zoo dataset demonstrate the accuracy and stability of the ECCC algorithm.

B. Scalability Test

The purpose of this experiment is to test the scalability of the ECCC algorithm when the sizes of the datasets increase. We picked the first 1K, 2K, 3K, 4K, 5K, 6K, 7K and 8K records respectively from the mushroom dataset to form 8 testing datasets. Figure 1 shows the run time of ECCC testing on the 8 datasets.

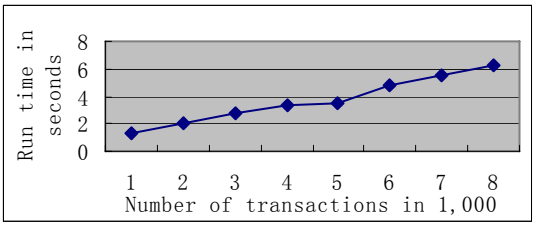


Figure 1. time vs. number of transactions

From Figure 1, it is easy to see that the computation cost increases linearly in terms of the number of transactions, which is highly desired in real data mining applications.

IV. CONCLUSION

In this paper, we gave a better quality index (especially the *Discriminativeness Index)* based on equivalence classes of frequent patterns, and proposed an equivalence class based clustering algorithm (ECCC) for categorical data. ECCC mines clusters with high intra-cluster similarity, strong inter-cluster diversity and no cross-cluster overlap. The experiment results showed that our method is accurate, stable and scalable on the real datasets.

For future work, we will find a good way to merge some clusters for reducing the number of clusters according to user's requirement.

ACKNOWLEDGMENT

Thanks go to the authors of [9] for the C++ code of DPMiner.

REFERENCES

[1] Weining Qian and Aoying Zhou. Analyzing Popular Clustering Algorithms from Different Viewpoints. Journal of Software, 13(8): 1382-1394, 2002.

[2] Z. He, X. Xu and S. Deng. K-ANMI: A mutual information based clustering algorithm for categorical data. Information Fusion, 9:223-233, 2008.

[3] Z. He, X. Xu and S. Deng. Squeezer: an efficient algorithm for clustering categorical data. Journal of Computer Science & Technology, 17(5): 611-624, 2002.

[4] D. Cristofor and D. Simovici. Finding median partitions using information-theoretical-based genetic algorithms. Journal of Universal Computer Science, 8(2): 153-172, 2002.

[5] Z. He, X. Xu and S. Deng. A cluster ensemble method for clustering categorical data. Information Fusion, 6(2): 143-151, 2005.

[6] T. Li, S. Ma and M. Ogihara. Entropy-based criterion in categorical clustering. ICML, 2004.

[7] Nicolas Durand and Bruno Cr´emilleux. ECCLAT: a New Approach of Clusters Discovery in Categorical Data. 22nd SGAI International Conference on Knowledge Based Systems, December 2002.

[8] http://archive.ics.uci.edu/ml/datasets.html, July 2011.

[9] J. Li, G. Liu and L. Wong. Mining Statistically Important Equivalence Classes and Delta-Discriminative Emerging Patterns. KDD, 2007.

[10] Qingbao Liu and Guozhu Dong. A Contrast Pattern based Clustering Quality Index for Categorical Data. IEEE ICDM, 2009.

[11] G. McLachlan and T.Krishnan. The EM algorithm and extensions. Journal of Classification, 15(1): 154-156, 1997.

[12] www.cs.waikato.ac.nz/ml/weka/, August 2010.

 ISBN: 978-1-61208-162-5

Exploiting Student Intervention System Using Data Mining

Samia Oussena, Hyensook Kim
University of West London
London, UK
samia.oussena@uwl.ac.uk, Hyensook.kim@uwl.ac.uk

Tony Clark
Middlesex University
London, UK
Tony.clark@mdx.ac.uk

***Abstract*—With the proliferation of systems that are put for the student use, data related to activities undertaken by the student are on the increasing. However, these vast amounts of data on student and courses are not integrated and could therefore not easily queried or mined. Therefore, relatively little data is turned into knowledge that can be used by the institution learning. In the work presented here, different data sources such as student record system, virtual learning system are integrated and analysed with the intention of linking behaviour pattern to academic histories and other recorded information. These patterns built into data mining models can then be used to predict individual performance with high accuracy. The question addressed in the paper is: how can indicators of problems related to student retention produced by data mining be presented in a way that will be effective. A prototype system that integrates data mining with an intervention system based on game metaphor has been build and piloted in the computing school. Early evaluations of the system have shown that it has been well received at all levels of the institution and by the students.**

Keywords-data mining; intervention system; student drop-out; game metaphore

I. INTRODUCTION

Whilst student engagement is complex and multi-dimensional, one key aspect for high education institutions is to engage students at a personalised level. There is a proliferation of data related to student activities. Activities might relate to some actions the student has performed such as submitting an assignment or viewing lecture notes. However, these vast amounts of data on student and courses are not integrated and could therefore not easily be exploited. Consequently, little data is turned into knowledge that can be used by the institution learning. Data mining is the field of discovering of implicit and interesting patterns for large data collections [8]. Data mining has been applied to a number of fields including bioinformatics and fraud detection. In recent years, there has been an increased interest in the use of data mining to the educational setting. Data mining has been shown to help predict student educational outcomes [13]. When using Data Mining, the goal is to develop a model, which can infer an aspect of the student academic outcomes, such as passing a module, from a combination of other data that represents student's characteristics. For example, Garbrilson [6] uses data mining prediction techniques to identify the most effective factors in determining student test scores. In [10], authors use a data mining classification technique to predict student's final grades based on their web use. In most of these applications, the results are usually presented to strategic decision makers in graphical form (for example a dashboard), which is interpreted and some intervention programme is then put in place. However the integration of the predictions provided by Data Mining, the presentation of the results produced by applying predictive patterns to live student data and the intervention processes are typically weak. In most cases intervention is provided by humans resulting in an overall process can be very resource intensive. There is therefore a lack of integration between the identification of the problem and the implementation of the intervention actions.

In this paper, we discuss the design of an intervention system based on data mining. We have developed an application that allows the data mining models to be refined as a consequence of intervention actions, as well as involving students in the process. There is evidence that personalising and signposting educational 'moments' contributes to a better learning environment [7]. Although the literature on retention points to the complexity of factors influencing retention, there is evidence that linking social and academic experience, and tailoring the learning environment to individual needs increases an institution's chances of retaining its students [1].

The challenge in designing the application for the student retention has been how to maximise student engagement. Arguably the weakest link in the process arises in ensuring that students at risk are identified as early as possible. Of course, this can be achieved with unlimited resources in the form of tutors who continually monitor raw data sources and who contact students as soon as they detect a problem indicator. However, this is not realistic. Our proposal is to automate the process and to present information to students in a way that will maximise their engagement and therefore reduce the resource burden. Hence our research question is: how can indicators of problems related to student retention produced by data mining be presented to students in a way that will be effective without an unrealistic resource overhead?

The strong widespread appeal of computer and console gaming has motivated a number of researchers to harness the educational potential of gaming [3]. Here, we have looked at using the motivation power of games to encourage students to be involved with the intervention system. Our hypothesis is that: features used by gaming systems can be incorporated into student intervention systems in order to maximise their effectiveness.

 ISBN: 978-1-61208-162-5

To test out hypothesis we have designed a prototype system that has the following features: we have designed a uniform data model describing a student profile within a teaching and learning environment; data mining techniques are then used to process the information and to produce rules that represent indicators of failure within the educational process; the rules are then processed against live student data in order to raise potential indicators of failure in real-time; gaming systems have been analysed in order to produce a model whereby information can be presented to students as though their learning experience is a game; this model has been implemented in the form of a web application and the events produced by the rules are fed into the gaming model.

The rest of the paper is as follows: Section 2 discusses some of the student retention work; Section 3 discusses the students and the learning models that we have used; Section 4 discusses our gaming environments; Section 5 discusses the design of the application.

II. BACKGROUND

Several modelling methods have been applied in educational research to predict student's retention. The more widely used models are the Students attrition model (Bean 1980) and the Tinto student integration model [17]. Tinto's model examines factors contributing to a student's decision about whether to continue their higher education. It claims that the decision to persist or drop out is quite strongly predicted by their degree of academic integration, and social integration. Tinto argues that from an academic perspective, performance, personal development, academic self-esteem, enjoyment of subjects, identification with academic norms, and one's role as a student all contribute to a student's overall sense of integration into the university.

Students who are highly integrated academically are more likely to persist and complete their degrees. The same is true from a social perspective. Students, who have more friends at their university, have more personal contact with academics, enjoy being at the university, and are more likely to make the decision to persist. Bean's model appears to use many of the constructs in Tinto's model but the most significant addition is the inclusion of external factors. These include attitude constructs which might have a direct effect such as finance or indirect such as influence of parents and friends' encouragements [2].

There are also other models of student retention. Thomas developed her model "institutional habitus" [18]), based on Tinto's theory, which can be divided into the academic and the social experiences. The academic experience covers attitudes of staff, teaching, learning and assessment. Different learning styles are supported and diverse backgrounds are appreciated. Tutors are friendly, helpful and accessible. Assessment gives students the opportunity to succeed and staffs are available to help. The social experience combines friendship, mutual support and social networks. Thomas noted that one factor in her students' persistence was the fact they felt more at home with their friends.

Recently, data mining models have also been developed for addressing student retention. For example, in [10], the authors use data mining classification techniques to predict students final grades based on their web-use feature. It can identify students at risk early and allow the tutor to provide appropriate advice in a timely manner. Cerrito applied data mining in mathematics courses, her study demonstrates that retention needs to be of concern at all levels of a student's career at the university, not just for the first year students [4]. Students with high entrance scores and low risk factors may also leave the university before graduation. However, most of these projects are lacking the following up intervention. Seidman's has shown that early identification of students at risk as well as maintaining intensive and continuous intervention is the key to increasing student retention [14]. He also explains how universities can prepare their programs and courses so students will have the greatest probability of success both personally and academically.

In this paper, we argue that the intervention process needs to be integrated with the identification process in order to be effective. The data mining process will help with the early identification, followed by early and continuous intervention, as proposed by Seidman. By combining the two processes, we are able to provide an audit trail of the intervention actions that can be then evaluated for their effectiveness.

III. AN INTERVENTION SYSTEM BASED ON GAMING ENVIRONMENT

Gaming, and particularly on-line gaming has become very popular with young people in recent years. In such systems, players can develop a profile based on playing a number of (possibly collaborative) games. The profile can be tailored to the individual in terms of the look and feel and represents achievements in terms of goals attained, points achieved, extra features unlocked etc. When a game is played, there are a number of features that can be attained such as completing a level or defeating a foe. In general, each game attaches points to the different challenges it presents and, although the games are different, the points are in a universal currency (or at least universal to a specific gaming platform). Points awarded to a specific gamer represents their level of achievement in terms of skills attained and challenges overcome. Gamers can compare their aggregate performance against other gamers to produce a league table; relative positions in a league table can be a powerful motivating factor and gamers can spend a great deal of time trying to move their position up the table. In addition, games can compare their performance at a more fine grain level in terms of specific skills and achievements. A gamer may have a specific interest in achieving a given skill because it is transferrable to another game.

Our proposal is that students can be viewed as gamers and learning outcomes can be viewed as being similar to points awarded when playing games. That being the case, we propose that the same powerful motivating factors that lead gamers to strive to increase their performance (whether relative or absolute) can be applied to students. A number of attributes common to computer games are recognized in fostering active engagement; motivation and a high level of persistence in game play [6]. These include the use of

 ISBN: 978-1-61208-162-5

environment that simulates realistic experiences for the player, providing opportunities for identity exploration and play through role play [15]. In a number of games, a player may learn to take on attributes of their avatar [19]. Using avatars may lead to a sense of responsibility towards the character that can lead to educationally relevant outcomes. The other main attribute is the creation of a sense of pride and accomplishment by structuring the game to challenge the player and allow progress.

The above principals have been implemented in the intervention system in order to harness the motivation potential of gaming; including associating academic performance with scores, use of avatar, structure the levels based on learning outcomes and modules and providing a league board.

IV. THE INTERVENTION SYSTEM OVERVIEW

We have built an intervention system that put students as the main actors. Students play a critical role in being successful and subsequently remaining at the university. Studies have indicated that motivation is a prerequisite for student learning [16]. The student can foster this motivation by setting clear and explicit learning goals and understanding the expectation of success. The greater the belief that a task can be accomplished the greater the motivation. In our system, at any point of time the student will be presented with what has been accomplished so far in terms of learning outcomes gained within each module, the modules that have been passed and the marks gained so far. The student is also aware of what is expected in order to acquire his qualification, for example, in terms of modules to be taken, learning outcomes to be gained within a module, and number of assignments. However, this will only make an impact if students are engaged i.e. access and make use of the information. Hence, while modeling the student we looked at both the information predicting their performance as well as information related to their engagement with the system.

The design of the predictive model part is based on two main information channels, as suggested by Tinto, one related to the student and one related to the university. The information related to university is based mainly on information related to courses and their related modules. The design of this part of the model has been constrained by the institution's information that we had access to. The constraints were mainly due to the interpretation of the data protection act. One of the techniques that supported our design was feature importance technique. Here, we have put all the data that we had access to and carried a data mining analysis to help us determine which indicators provide the best assessment of potential student academic performance. As illustrated in Figure 1, the model includes the following types of information:

- Information related to their entry profile such as their qualification and the scores of the tests that they took at their entrance.

- Information on the course that they enrolled on. This will contain a general information related to the course itself such as the modules that are part of the course, the faculty that manages the course, the type of award, and a more specific information related to the student enrolment to the course offering; for example, the year the student has enrolled and how many modules he had to repeat in that course.

- Information related to individual module that the student is enrolled on. Here also it includes a general information related to the module such as the number of assignment for the module, number of credits for the module and the team delivering the module, and information related to the student enrolment to the specific module offering such as the date enrolled and the results for that module.

- The other type of information relates to the interactions the student has with the module resources; such as interaction with the virtual learning site, interaction with library, interaction related to the module assessment such as submission of assignment, results acquired for the assignment.

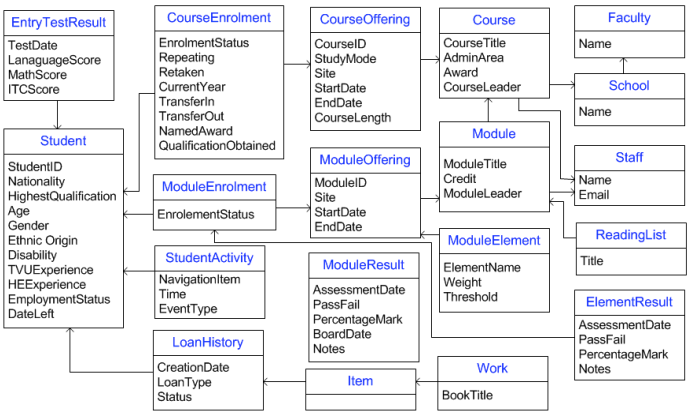


Figure 1. Student model: predictive part

The design of the engagement model part is based on the mapping of student performance to the game attributes. For example, based on their profile, students are associated to one of the groups (Zone) that have been identified by the data mining process. For each of the zone, the data mining process would have identified a threshold for when intervention actions are required. In this model, we have included not only individual performance information but also performance information related to the group, the module and the course.

 ISBN: 978-1-61208-162-5

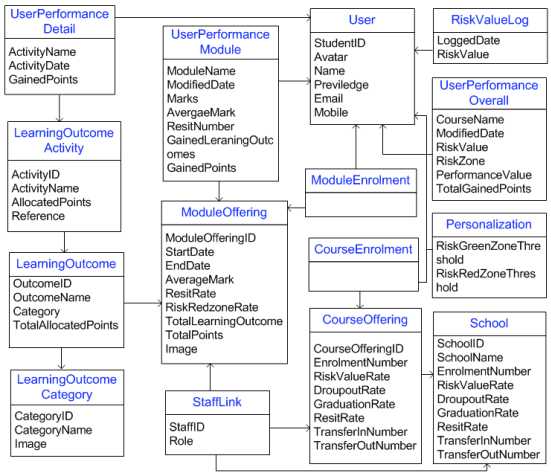


Figure 2. Student model: engagement part

As illustrated in Figure 2, the types of information included in the model are the following:

- Personal information such as their Avatar, their preferred mode of communication.
- Information related to points that they have acquired and need to acquire.
- Information related to module; these include all the learning outcomes that are associated to it, the activity that allows the achievement of the learning outcome and the number of points associated to it. They will get predefined points for doing activities such as accessing blackboard, borrowing a book from library, or submitting an assignment. The blackboard visits in a day for one particular module will be counted as 1 visit. If the student visits blackboard to access resources for another module on the same day, another visit will be added related to that module. Similarly they will obtain points when they obtain marks for each of the assessment. They will also unlock learning outcomes for the modules when they do the activities related to the learning outcome. They will unlock all the learning outcomes for a module when they pass the module.

The architecture of the system is illustrated in Figure 3. The system has been built using Oracle technology [11] and includes three main components; the data warehouse component, the data-mining component and the intervention application. We have used oracle workflow system to execute data warehouse update and run data mining engine. Currently the workflow runs every Friday. The intervention reporting system runs on its own scheduler after data mining process.

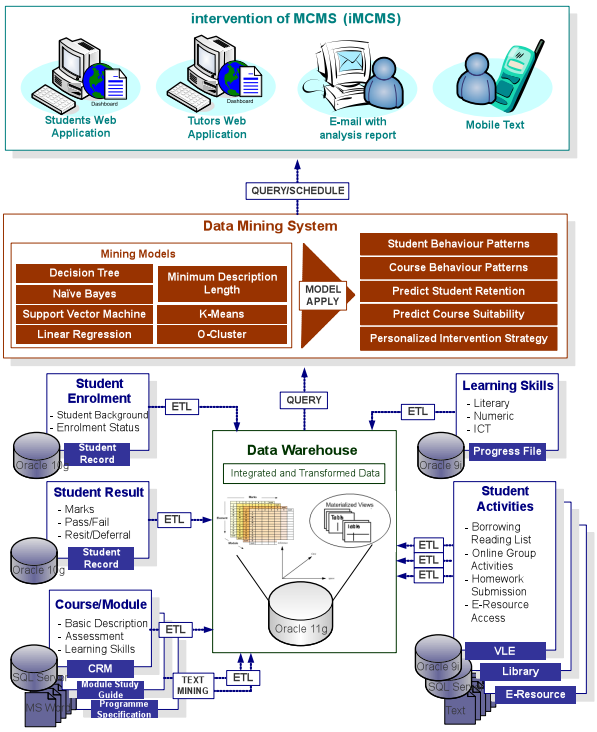


Figure 3. MCMS System Architecture

A. *The data warehouse component*

The data warehouse component takes the data provided by the institution data sources and builds the data warehouse; i.e. the different dimensions of the cubes. The data sources used for the system are discussed in the next section. We have defined four cubes in this study; one for students, one for student activities, one for modules and the other one for courses.

B. *Data sources*

When analyzing the data sources, we have identified information that will give us a good understanding of the students' profile; i.e. data related to information before their entry to the university (student background), data related to their interaction in the university; including their goals (student interaction), and data related to their results. The key sources used are as follows:

- The **student record system** relates to the profile of the student prior to joining the institution. Student profile will include information such as their entrance level, their ethnicity, literacy and numeracy test entrance score. In addition the student's assessment profile including marks and the number of re-sits taken is maintained by the record system.
- The **library system** and the **reading list system** captures information about a student's book loan activities and links this to the reading lists set for the modules that the student is registered for.

- The **online learning system** and **e-library** records how often the student logs into the system and the use of the various pages in the *virtual learning environment* (VLE). In our case the VLE can capture the number of hits on individual pages and document downloads.
- The **module study guides** provide information such as reading lists and the schedule of assessments for each module. The **course specification** provides information such as regulations about options and assessment hurdles. The study guides and course specifications also provide lists of learning outcomes that can be gained by the students when they pass individual modules or module elements. Individual assessments are broken down into different types. Institution departments that own course components are recorded in addition to the tutor responsible for delivering the module.
- The **marketing system** and **entrance test system** provides information about entry requirements and whether students have had any additional tutoring on entry. For example overseas students may be offered extra tuition in English and the test system will record whether they took up this offer and, if so, the results.

C. *The data mining component*

Educational data mining is a newly emerging discipline and there have been reports of a number of demonstration applications in Spanish [13] and American [9] universities, and particularly in distance-learning institutions [11]. Data Mining has already proved to be successful in e-commerce and bio-informatics, where results are achieved through the use of associators, classifiers, clusterers, pattern analysers, and statistical tools. In the educational context, data mining provides analysis of the students' behaviour, navigation, frequency, and length of interaction with the e-Learning system that can identify patterns of behaviour and associations. It can classify students into groups depending on their learning behaviour rather than just ability. At the same time it also identifies students exhibiting atypical behaviour that needs early intervention and feedback. The overall process that we have used in this system is illustrated below. At each iteration, data is fed to the data mining process, in order to identify potential drop-out students and their performance. The intervention process will then identify specific intervention actions for these students.

The data mining process that we have adopted includes three main stages: finding the features' relations, data grouping, and making the prediction [20] The first stage involves applying features' importance and associate rules to find the correlation among data features. This stage helps eliminate any data that is unlikely to make any impact on subsequent tasks. For example, we found that age and gender is not related to students' performance (results or drop-out), whereas the VLE interaction is related to students' performance.

The next stage involves classification of data and extraction rules and patterns. Here we identify groups of students that have shown related features and will require similar intervention. Examples of such groups include first year full time students, students transferring from other institutions, non-UK students, students with low library usage or post-graduate students. The third and final stage includes applying regression to the identified groups in order to predict future behaviours i.e. allowing us to predict the potential students that are at risk of dropping out, or their academic performance requires attention. Here we identify students that are underperforming or the ones that have shown an improvement in their performance.

The data mining models have been built based on three years of historical data before being integrated into live feed data. The data were divided in two sets, one for building the models and one for validating the model. Once the models have been validated, the data warehouse is updated every week. The prediction stage of the process is conducted on the updated data leading to new alerts and update of the data.

D. *Intervention Application Component*

This component presents the results of the data mining models and implements the intervention system. We have used different views for different stakeholders. One view is targeted towards improving students' performance and providing them with a holistic and a detailed view of their performance. Students at risk of dropping out receive an intervention message via email or SMS. The other view is to support the academics in implementing their intervention policies. The data here is presented at different levels of abstractions depending on their role (tutor, programme leader, and head of school). Any of the intervention content is generated according to a predefined and personalised intervention rule.

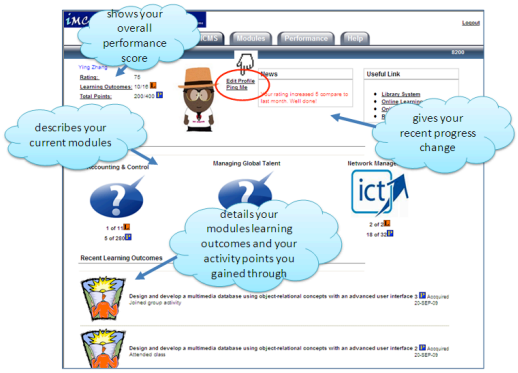


Figure 4. Student Intervention Application

Students are encouraged to use avatars that have characters that they would like to relate to. Being associated with a strong avatar might motivate them to acquire high scores. Student activities are mapped to module learning outcomes activities in order to convert student performance as counted points. For example, if a student borrowing a book from the reading list, he or she would get points for his

or her module and if he or she submits an assignment, he or she gets points as well. If the students reach the target points for each learning outcome of a module, they would see their mission achievement records in the web application and would encounter the next challenge for further improvement such as a usual role-playing game. We have also implemented a league table where scores can be compared within the same group (Zone). A screenshot of the implemented student web application is shown in Figure 4:

Students are alerted by email or text if they are in at-risk zone. In the message, they are presented with their achievements so far in terms of the points. The message also advises them with a number of actions that they need to undertake in order to get out of the zone. A student that has taken the right action and has managed to get out of the zone is also alerted. The message that they receive is a congratulatory message that invites them to review their achievements and show them how they can unlock the next level.

The design of the messages and the actions that the students are invited to undertake are associated with each of the groups that were identified in the second stage of the data mining process. For each group, based on the pattern that describes the group, a template for the message is designed, with specific actions that will lead to correcting the low scores in the particular interaction, such as accessing the module VLE content. The templates are then personalised with live data specific to each student. The intervention messages are common to all modules, however each module leader is encouraged to refine the rules that will lead to alerts and customise the intervention messages and actions.

V. DISCUSSION AND CONCLUSION

Through this paper, we have explored the benefitting factors of using a gaming environment in order to engage and motivate students in terms of retention. The prototype system has been used to integrate the data mining process with the intervention system. The data mining process is used as two main functions; primarily to predict those students who are most likely to drop-out early, and secondly to group students into specific categories that can be targeted with personalised interventions if it is predicted that a drop-out is imminent. Certain activities and accomplishments merit a specific number of points to be added on the appropriate student's game profile. The student unlocks each level in the gaming process through achieving a set score for each learning outcome. This idea of levels is a great source of motivation to students, as well as the idea of a 'leader board' in which students' are encouraged to become competitive. This also enforces the gaming environment, making 'winning' appear more appealing than in a classic university environment.

The system has been piloted for a semester in the computing school and the institution is planning to pilot it in more schools in the next academic year. The early evaluation of the system has mainly been done through presentation and focus groups. The questions that we tried to address in these focus groups are for example; how do student perceive the intervention application impact on learning and their academic performance? Are the predictions of the data-mining models accurate? Is the gaming metaphor a good motivation tool? How do tutors and university staff perceive the impact of the intervention application on the student learning and academic performance? What improvements were made to some of the learning and teaching processes? The early results have shown that the system has been well received and the prediction have been very similar to what the tutor expected. We are planning to do a more in depth quantitative data analysis in order to track student engagement and survey their perceptions of the environment.

In terms of further improving the system there are a few problems that must be targeted. For example, in each iteration, the data mining models must be manually retrained. This is a hindrance and can be rectified by implementing an algorithm that creates self-adapting models. Another area to expand upon is the range of data sources available. We hope to add new data sources, such as financial data, timetabling and data relating to their social involvement in universities. This has not initially been possible due to data constraints. Currently, another part of the system that has not been discussed in this paper, involves the monitoring of courses and modules. This has given us some insight into the performance of the education processes and one of our planned works is to investigate the relationship between the performance of the education processes and the student performance. The other area for improvement will involve creating an additional gaming feature (i.e. settings), providing a user-friendly interface enabling students to access key information, such as course information, university regulations, and their personal timetable. This idea of integrating learning and games could be further developed through literacy and numeracy features to support their main subject. This is once again could be incorporated into the gaming environment, similar to many educational games and consoles available in the current market.

REFERENCES

[1] K. Anagnostopoulou, and D. Parmar, Practical guide: bringing together e-learning & student retention, London: Cats Ltd. ISBN: 978-1-85924-301-5.

[2] J. Bean, Dropouts and turnover: The synthesis of a causal model of student attrition. Research in Higher Education, 12, pp. 155-187.

[3] S. DeCastell and J. Jenson, Serious play, Journal of curriculm Studies 35(6), pp. 649-665

[4] P. Cerrito, Data Mining Student Performance in Mathematics Courses, CinSUG Third AnnualOne Day Conference. Committee of Public Accounts, (2002) Fifty-eighth Report of Session Improving Student Achievement and Widening Participation in Higher Education in England, HC 588,2001-02.

[5] S. Gabrilson, Data Mining with CRCT Scores. Office of information technology, Geargia Department of education.

[6] R. Garris, R. Ahlers, and J. Driskell, Games, motivation, and learning: A research and practice model. Simulation and Gaming: An Interdisciplinary Journal, 33(4), pp. 441–467.

 ISBN: 978-1-61208-162-5

[7] L. Harvey, S. Drew, and M. Smith, The First-year Experience: A Review of Literature for the Higher Education Academy, http://www.heacademy.ac.uk/research/Harvey_Drew_Smith.pdf

[8] W. Kloesgen and Zytkow, Handbook of Knowledge Discovery and Data Mining, Oxford University Press, Oxford, 2002.

[9] J. Luan, Data mining and knowledge management in higher education –potential applications. In Proceedings of AIR Forum, Toronto, Canada.

[10] B. Minaei-Bidgoli, G. Kortemeyer and W. Punch, Enhancing Online Learning Performance: An Application of Data Mining Methods, From Proceeding of Computers and Advanced Technology in Education .

[11] E. Mor and J. Minguillón, E-learning personalization based on itineraries and long-term navigational behavior, Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters,pp. 264-265 May 19-21, New York, NY, USA .

[12] Oracle http://www.oracle.com/technology/products/bi/odm/index.html [accessed on 25/05/2011]

[13] C. Romero and S. Ventura, Data mining in e-learning, Southampton, UK: Wit Press.

[14] A. Seidman, Retention Revisited: RET = E Id + (E + I + C)Iv. College and University, 71(4), pp-18-20.

[15] K. Squire, H. Jenkins, W. Holland, H. Miller, A. O'Driscoll and K. Tan, Design principles of next-generation digital gaming for education. Educational Technology, 700 33, pp.17–23.

[16] M. Svinicki, A. Hagen and D. Meyer, How to Reaserch on Learning Strengthens instruction, Teaching on Solid Ground: Using Scholarship to Improve Practice, Jossey-Bass publishers, San Francisco, CA.

[17] V. Tinto, Dropout from Higher Education: A Theoretical Synthesis of Recent Research, Review of Educational Research vol.45, pp. 89-125,1975.

[18] L. Thomas, Student retention in higher education: the role of institutional habitus", Journal of Education Policy, Vol. 17 No. 4, August, pp. 423-442, 2002.

[19] N. Yee, and J. Bailenson, The Proteus effect: The effect of transformed self-representation on behavior. Human Communication Research 33, pp. 271-290.

[20] Y Zhang, S. Oussena, T. Clark. and H. Kim, Use Data Mining To Improve Student Retention in Higher Education – A Case Study, to be presented in 12th International Conference on Enterprise Information Systems(ICEIS) pp. 190-197.

 ISBN: 978-1-61208-162-5

Supporting Global Design Through Data Mining and Localization

Barbara Rita Barricelli
Department of Computer Science and Communication
Università degli Studi di Milano
Milan, Italy
barricelli@dico.unimi.it

Malte Ressin
Centre for Internationalization and Usability
University of West London
London, UK
malte.ressin@uwl.ac.uk

Abstract **Localization can be an important work step for software development with a considerable impact not only on success, but also on cost and quality. To facilitate localization, a number of tools exist. In particular in the area of computer-assisted translation, data analysis is used to aid in the work of the translator. In this paper, we propose to apply data mining to assist in the localization of software elements beyond text translation, such as colors, symbols and images. In particular, we propose to apply data mining to make the most of available resources on the internet and treat them as distributed databases.**

Keywords-global design; localization; data mining; translation; internationalization; globalization; culture

I. INTRODUCTION

The World Wide Web plays the fundamental role of medium for international communication, participation, and transaction. The characteristics of the Web, its tools and technologies support and stimulate the evolution of methods and techniques for interface design for multi-cultured environments [1]. However, in order to guarantee an international usability, the software applications have to be designed in a culture-oriented way [2][3].

To design and develop global products means in fact to a) extend it to different international contexts, b) to make it able to handle various languages and conventions, c) to localize it according to specific cultures, and d) to translate it in the proper languages.

Up to now many efforts have been made in the field of machine-based translation, and data mining techniques are widely used to this end. However, other aspects related to cultures that are not related to the languages are not yet taken into account.

The contribution of this paper is twofold. First, we highlight the challenges that emerge from the current asset of Web, its global access and the spread of its technologies. Second, we propose a data mining application for the localization of software applications which is able to make the most of available resources on the internet and treat them as distributed databases.

This paper is organized as follows. Section II illustrates the theoretical and practical backgrounds in the field of global design and localization. Section III presents a review of the current tools of Computer-Assisted Translation. In section IV, the challenges that arise from global software design are highlighted and a proposal about the implementation of data mining tools is presented. Section V closes the paper proposing open questions to be addressed in future development of this research.

II. GLOBAL DESIGN AND LOCALIZATION

In the global software design literature, many definitions of culture have been given. Yeo [5] defined it as "behavior typical of a group or class (of people)"; Bødker and Pedersen [6] defined culture as "a system of meaning that underlies routine and behavior in everyday working life"; while for Borgman [7], culture includes "race and ethnicity as well as other variables and is manifested in customary behaviors, assumptions and values, patterns of thinking and communicative style".

The literature about culture and its meaning shows that for a long time, this topic has been discussed in the field of computer science. However, the outcomes of these discussions are still limited and have not been applied concretely and completely to software design and development. Our effort in this research area is to study, propose and develop methods and techniques able to respond to the challenges that emerge from the global design context in semi-automatic ways. We consider it in fact necessary to involve experts in localization to validate the results of automatic tools, because the human knowledge and expertise has to be reinforced and not replaced. Our work stems from the study of the literature about culture, cultural dimensions and localization methodologies and best practices that already exist and that we describe in what follows.

A. Cultural Dimensions

In literature, various cultural models have been proposed and each of them is described by a set of cultural dimensions [8][9][10][11][12]. The most adopted and discussed classification of cultural dimensions is the one proposed by Hofstede [12], by which he recognized five main dimensions:

- Small vs. large power distance: measures the extent to which the less powerful members of organizations and institutions accept and expect that power is distributed unequally.
- Individualism vs. collectivism: measures the degree to which members of organization and institutions are integrated into groups.
- Masculinity vs. femininity: measures the distribution of roles between the genders.

 ISBN: 978-1-61208-162-5

- Weak vs. strong uncertainty avoidance: measures to what extent a culture prepares its members to feel either uncomfortable or comfortable in unstructured situations. Uncertainty-avoiding cultures try to minimize the possibility of unknown and surprising situations by strict laws and rules, safety and security measures, and on the philosophical and religious level by a belief in absolute Truth. Uncertainty-accepting cultures are more tolerant of opinions different from what they are used to; they try to have as few rules as possible, and on the philosophical and religious level they are relativist and allow many currents to flow side by side.
- Long vs. short term orientation: measures to what extent a culture respects values associated with long term orientation or short term orientation. Long term orientation values are thrift and perseverance, while values associated with short term orientation are respect for tradition, fulfilling social obligations.

Several studies have shown how these five cultural dimensions, which classify a person's cultural background into certain scores, relate to certain aspects of a user interface [13][14][15]. One of the most interesting categorizations is the one given by Yeo [5] that categorizes the factors needed to be addressed in global design processes into covert and overt. Overt factors are tangible, straight forward and publicly observable elements. Some examples are date, calendars, time, address formats, character sets, punctuation, and currency. Covert factors are those elements that are intangible and culture-dependent. Colors, sounds, metaphors are examples of covert factors.

B. GILT Methodological Model and Best Practices

The design and the development process of global products passes through the performance of four distinct activities [16]: globalization, internationalization, localization, and translation. These four activities constitute the so-called GILT methodological model. Internationalization is an activity that is performed independently by localization and translation, because it affects the structure of the product under design and development and not its content. Translation is included in localization because it represents just one of the actions required to localize a product. An example of internationalized and localized Website is Wikipedia, as shown in Figure 1.

The localization activity is detailed in [17], by separating it into two distinct components, content and package. Content is defined as the linguistic structures, while package is the set of all the non-textual elements and the media through which the content is distributed. After Esselink [4], globalization "addresses the business issues associated with taking a product global. In the globalization of high-tech products this involves integrating localization throughout a company, after proper internationalization and product design, as well as marketing, sales, and support in the world market". Globalizing means therefore the extension of a product to different international context, with the aim of making it usable by the different potential users.

Figure 1. Two localizations of Wikipedia: (a) English and (b) Arabic. The organization of the page follows the writing direction of the language (from left to right for English and from right to left for Arabic).

Internationalization is "the process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for re-design. Internationalization takes place at the level of program design and document development". Localization "involves taking a product and making it linguistically and culturally appropriate to the target locale (country/region and language) where it will be used and sold", while translation is "only one of the activities in localization; in addition to translation, a localization project includes many other tasks such as project management, software engineering, testing, and desktop publishing". While translation is aimed at maintaining the meaning of original information by exposing them in different languages, localization transforms the information in equivalent ones but adapted to a different culture.

As suggested by [18], in order to develop software suitable for the global market, a two-step process is needed: internationalization of the software first and its localization next. In [19], a cross-cultural checklist that should be considered by interface designers is given. The authors consider several key factors:

- Text: a simple translation is not enough, many aspects should be taken into account (e.g. jargon, character sets, numbers, date, time formats).
- Images: images represent the visual language of a culture and therefore not only image recognition but also image acceptability problems should be considered.
- Symbols: as for images, also symbols have to be acceptable for the target culture.
- Colors: as pointed out in [20][21][22] interpretation of colors varies in the various cultures. Colors play a

fundamental role in interface design because they convey information, and therefore they need to be chosen very carefully.
- Flow: the writing system of a language and therefore its reading/writing direction affects the way in which the information is recognized by users on a screen. Hence, the logical flow of what is represented on an interface should follow the proper directions.
- Functionality: sometimes functionalities implemented in software application are not accepted in some cultures because they do not respect the cultural conventions that the user needs.

III. CURRENT SOFTWARE TOOLS

A number of software tools are available to assist translators in their work. These tools are exclusively aimed at text translation. Their use is commonly called Computer-Assisted Translation (CAT), the software suites incorporating them are usually referred to as Translators' Workbenches. Most of these software tools are centered on extended data collection and database searches and fall into the following two rough categories:

- Translation Memories (TMs), also called repetition manager, store previously translated content and keep it accessible for future use by the translator.
- Machine Translation (MT) employs computer algorithms to derive translations.

A. Translation Memory

Nowadays, the use of TMs is widespread throughout commercial translation [23][24]. A number of commercial as well as free translator workbenches are available [25]. Often, the TMs in these come with tools which automatically scan the text of source documents and provide close matches found in its database. Provision of such suggestions can save the translator precious time not having to look up previous translations, while at the same time increasing translation consistency. Accordingly, Schäler [26] has found that usage of such tools provides the following benefits:

- Speed up the translation process.
- Improve translation quality.
- Reduce translation cost.

However, Ottman [27] argues that the use of TMs introduces an additional source of translation errors through incorrect context. As a consequence, she argues that additional quality assurance might be necessary.

B. Machine Translation

Broadly speaking, there are two approaches to MT. Rule-Based Machine Translation (RBMT) aims to translate text through the use of dictionary and grammar encoded as a program. Statistical Machine Translation (SMT) compares the source text with existing bilingual text to derive a translation of the source text. Combinations of both approaches exist.

For the context of this article, only SMT is of interest. This approach relies on analysis of text which is already available in different languages, similar to the Rosetta stone. The results of this analysis are then used to translate new source texts.

The potential use of MT as tool for translation has been understood early on [28][29]. However, pure machine translation still requires extensive human review due to prevalent quality issues [30]. Elsen [31] states that MT has the potential to increase translation speed while at the same time reducing translation cost. Although he asserts that MT should be particularly applicable to short and simple text, a requirement satisfied by typical user interface text, he concludes that placeholders would pose additional difficulty. An example of MT tool is given in Figure 2.

Figure 2. Google translate, one of the most used MT tools available online.

IV. CHALLENGES IN DATA MINING AND LOCALIZATION

While the tools mentioned in the previous chapter can be employed for software localization, they will be of assistance only for text translation. The remaining visual elements described in chapter II are not covered by these tools.

From our experience in localization and internationalization and from the critical analysis of the literature, the need of semi-automatic tools to support the software developers in choosing the right visual elements is strongly emerging.

However, we consider the localization process as an activity that should be managed by an interdisciplinary team: developers are in fact not able to deal with cultural issues alone and they need to work together with translators, ethnographers, and other experts. This is the reason why we are not promoting the development of automatic systems but of semi-automatic systems that should be designed to help the team and not to substitute the human tasks. The experience and background of all the stakeholders in the localization team should be exploited and applied to the cases at hand. In the category of visual elements, we consider all the elements that could be part of an interface and are not textual, e.g., colors, pictures, symbols.

Ryan, Anastasiou and Cleary (2009) suggested use of a Localization Knowledge Repository (LKR) to facilitate localization of such elements.

The goal of this paper is to propose the adoption of data mining techniques to address the challenges that emerge from the state of the art in localization and internationalization. Data mining could be in fact the good means by which to derive from existing applications the rules to be used for the design and development of new ones.

 ISBN: 978-1-61208-162-5

We identified three main challenges that are related to the choice of colors, symbols, and images. A data mining tool could be used to make comparisons between the various localized instances of internationalized application in order to derive some cultural rules that regard the use of colors and the appropriate symbols and images to be used. For instance, by comparing two instances of the same Website, one localized for a culture and the other localized for a different one, we could observe the different choices made in order to deliver the same meanings but using different visual elements. Clearly, such kind of analysis could be significant only if applied to a large number of software applications. Another help could be given by the analysis of the meta-tags included in the code of the websites which could add some information about context and content of the application.

For our proposed application, we suggest the use knowledge discovery and known techniques of software mining on the code level. Specifically, software is mined for patterns at the user interface and data level, possibly also at the business code and statement level. Treatment of individual elements would differ on element complexity. For example, color can be normalized into RGB codes. More complex elements such as symbols and images would require additional processing, for example via image recognition. Similarly, mining internationalized software would differ marginally from mining non-internationalized software insofar as different releases or (language) versions might have to be processed in case of the latter. However, both could feed into the resulting data sets.

This software mining will enable the creation of association rules between software properties or elements marked through markups, meta-tags, and localization implementations. Additionally, it is conceivable to use surrounding code as marker. After normalizing those association rules, they can be applied to new, yet unlocalized applications.

V. CONCLUSIONS AND OPEN QUESTIONS

In this paper, we discussed software localization and its facilitation by software tools. We suggested leveraging data analysis to provide culture-conforming colors, symbols, images etc. already during the design phase.

The approach we suggested is obviously heavily influenced by statistical machine translation. As such, it shares its downsides, i.e. the requirement of identical documents for different languages or cultures. As we elaborated above, in order to be useful, these documents need to provide sufficient context information, for example via the use of meta-tags. Such explicit information can only be avoided if there is sufficient implicit data to ensure the appropriateness of an element's or item's translation in the given context.

We firmly believe that the potential of data analysis methods for localization has not yet been fully realized, and that there are a plethora of opportunities to further simplify localization beyond the realm of text translation.

ACKNOWLEDGMENT

The work of Barbara Rita Barricelli was supported by the Initial Training Network "Marie Curie Actions", funded by the FP 7-People Programme with reference PITN-GA-2008-215446 entitled "DESIRE: Creative Design for Innovation in Science and Technology".

REFERENCES

[1] W. Barber and A. N. Badre, "Culturability: The Merging of Culture and Usability," Proc. 4th Conference on Human Factors and the Web, June 1998.

[2] K. Reinecke and A. Bernstein, "Predicting user interface preferences of culturally ambiguous users," Proc. of 26th Conference on Human Factors in Computing Systems (CHI'08), ACM Press, Apr. 2008, pp. 3261-3266.

[3] E. M. del Galdo and J. Nielsen, International Users Interfaces. New York, NY: John Wiley & Sons, 1996.

[4] B. Esselink, B. A Practical Guide to Localization. Amsterdam, NL: John Benjamins Publishing Company, 2000.

[5] A. Yeo, "Cultural user interfaces: a silver lining in cultural diversity," SIGCHI Bull., vol. 28, no. 3, 1996, pp. 4-7.

[6] K. Bødker and J. Pedersen, "Workplace Cultures: Looking at Artifacts, Symbols and Practices," in Design at Work: Cooperative Design of Computer Systems, J. Greenbaum and M. Kyng, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991, pp. 121-136.

[7] C. L. Borgman, Cultural diversity in interface design. SIGCHI Bull., vol. 24, no. 4, 1992, p. 31.

[8] N. L. Hoft, "Developing a cultural model," in International Users interfaces, E. M. del Galdo and J. Nielsen, Eds., New York, NY: Wiley & Sons, 1996, pp. 41-73.

[9] E. Hall, The silent language. New York, NY: Doubleday, 1959.

[10] F. Trompenaars, Riding the waves of culture. London, UK: Nicholas Brealey publishing, 1993.

[11] D. Victor, International business communications. New York, NY: Harper Collins, 1992.

[12] G. Hofstede, Cultures and organisations: software of the mind. New York, NY: McGraw Hill, 1991.

[13] C. Dormann and C. Chisalita, "Cultural Values in Web Site Design," in Proc. 11th European Conference on Cognitive Ergonomics (ECCE11), Sep. 2002, pp. 8-11.

[14] A. Marcus, "Cultural Dimensions and Global Web Design: What? So What? Now What?," in Proc. 7th Conference on Human Factors and the Web, June 2001, pp. 1-15.

[15] A. Smith and Y. Chang, Y. "Quantifying Hofstede and Developing Cultural Fingerprints for Website Acceptability," in Proc. 5th International Workshop on Internationalisation of Products and Systems (IWIPS 2003), P&SI, July 2003, pp. 89-102.

[16] P. Cadieux and B. Esselink, "GILT: Globalization, Internationalization, Localization, Translation," LISA Globalization Insider, vol. 1, no. 5, 2002.

[17] M. O'Hagan and D. Ashworth, Translation-mediated communication in a digital world Facing the challenges of globalization and localization. Clevedon, UK: Multilingual Matters LTD, 2002

[18] T. Madell, C. Parson, and J. Abegg, Developing and localizing international software. Upper Saddle River, NJ: Prentice-Hall, Inc., 1994.

 ISBN: 978-1-61208-162-5

[19] P. Russo and S. Boor, "How fluent is your interface?: designing for international users," in Proc. Conference on Human Factors in Computing Systems (INTERCHI '93), IOS Press, Apr. 1993, pp. 342-347.

[20] L. G. Thorell and W. J. Smith, Using Computer Color Effectively: An Illustrated Reference. Englewood Cliffs, NJ: Prentice Hall, 1990.

[21] K. Garland, "The use of short term feedback in the preparation of technical and instructional illustration," in Proc. Conference on Research in Illustration, 1982.

[22] A. J. Courtney, "Chinese Population Stereotypes: Color Association," Human Factors, vol. 28, no. l, 1986, pp. 97-99.

[23] K.-H. Freigang and U. Reinke, "Translation-Memory-Systeme in der Softwarelokalisierung [Translation-Memory-Systems in software localization]," in Einführung in die Softwarelokalisierung [Introduction to Software Localization], D. Reineke and K.-D. Schmitz, Eds. Tübingen, Germany: Narr, 2005, pp. 55-71.

[24] E. Yuste, "Corporate Language Resources in Multilingual Content Creation, Maintenance and Leverage," in Proc. 2nd International Workshop on Language Resources for Translation Work Research and Training, Aug. 2004, pp. 9-15.

[25] S. Falcone, "Translation Aid Software - Four Translation Memory Programs Reviewed," Translation Journal, vol. 1, no. 2, 1998.

[26] R. Schäler, R. "A Practical Evaluation of an Integrated Translation Tool during a Large Scale Localisation Project," in Proc. 4th Conference on Applied Natural Language Processing (ANLC'94), 1994, pp. 192-193.

[27] A. Ottmann, "Lokalisierung von Softwareoberflächen [Localization of Software User Interfaces]," in Einführung in die Softwarelokalisierung [Introduction to Software Localization], D. Reineke and K.-D. Schmitz, Eds. Tübingen, Germany: Narr, 2005, pp. 101-115.

[28] C. Brace, "Trados: Ten Years On," Language Industry Monitor, July-August 1994.

[29] R. W. Collins, "Software Localization: Issues and Methods," in Proc. 9th European Conference on Information Systems, June 2001, pp. 36-44.

[30] J. Yao, M. Zhou, T. Zhao, H. Yu, and S. Li, "An Automatic Evaluation Method for Localization Oriented Lexicalised EBMT System," in Proc. 19th Conference on Computational Linguistics (COLING 2002), Aug. 2002.

[31] H. Elsen, "Maschinelle Übersetzung in der Softwarelokalisierung. [Machine Translation in Software Localization]," in Einführung in die Softwarelokalisierung [Introduction to Software Localization], D. Reineke and K.-D. Schmitz, Eds. Tübingen, Germany: Narr, 2005, pp. 89-99.

 ISBN: 978-1-61208-162-5

ArmSquare: An Association Rule Miner Based on Multidimensional Numbered Information Spaces

Iliya Mitov, Krassimira Ivanova
Institute of Mathematics and Informatics, BAS
Sofia, Bulgaria
mitov@mail.bg, kivanova@math.bas.bg

Benoit Depaire, Koen Vanhoof
Hasselt University
Hasselt, Belgium
benoit.depaire@uhasselt.be, koen.vanhoof@uhasselt.be

Abstract **– In this article, we propose a simple approach for association rule mining, which uses the possibilities of the multidimensional numbered information spaces as a storage structures. The main focus in the realization of ArmSquare is using the advantages of such spaces, i.e., the possibility to build growing space hierarchies of information elements, the great power for building interconnections between information elements stored in the information base, and the possibility to change searching with direct addressing in well structured tasks. The tested types of implementations of realized tool show the vividness of proposed approach.**

Keywords – Association Rule Mining; Market Basket Analysis; Multidimensional Numbered Information Spaces.

I. INTRODUCTION

Data mining stands at the crossroad of databases, artificial intelligence, and machine learning. Association rule mining (ARM) is a popular and well researched method for discovering interesting rules from large collections of data. Association rule mining has a wide range of applicability, such as market basket analysis, gene-expression data analysis, building statistical thesaurus from the text databases, finding web access patterns from web log files, discovering associated images from huge sized image databases, etc.

The contemporary databases are very large, reaching gigabytes and terabytes, and the trend shows further increase. Therefore, for finding association rules one requires efficient scalable algorithms that solve the problem in a reasonable time. The efficiency of frequent itemset mining algorithms is determined mainly by three factors: (1) the way candidates are generated; (2) the data structure that is used; and (3) the implementation details. Most papers focus on the first factor, some describe the underlying data structures, and implementation details are almost always neglected [1].

A. Problem description

The description of the problem of association rule mining is firstly presented in [2]. Below, the description of the problem follows one given in [3].

Let $\Im$ be a set of items. A set $X = \{i_1, ..., i_k\} \subseteq \Im$ is called an itemset or a k-itemset. A transaction over $\Im$ is a couple $T = (tid, I)$ where tid is the transaction identifier and I is an itemset. A transaction $T = (tid, I)$ is said to support an itemset $X \subseteq \Im$ if $X \subseteq I$. A transaction database D over $\Im$ is a set of transactions over $\Im$. The cover of an itemset X in D consists of the set of transaction identifiers of transactions in D that support X. The support of an itemset X in D is the number of transactions in the cover of X in D: $support(X, D) := |cover(X, D)|$. An itemset is called frequent if its support is no less than a given absolute minimal support threshold σ. The collection of frequent itemsets in D with respect to σ is denoted by $F(D, \sigma) := \{X \subseteq \Im \mid support(X, D) \geq \sigma\}$.

Problem 1. (*Itemset Mining*) Given a set of items $\Im$, a transaction database D over $\Im$, and minimal support threshold σ, find $F(D, \sigma)$.

An association rule is an expression of the form $X \Rightarrow Y$, where X and Y are itemsets, and $X \cap Y = \{\}$. X is called the body or antecedent, and Y is called the head or consequent of the rule. The support of an association rule $X \Rightarrow Y$ in D, is the support of $X \cup Y$ in D. An association rule is called frequent if its support exceeds a given minimal support threshold σ. The confidence of an association rule $X \Rightarrow Y$ in D is the conditional probability $P(Y \mid X)$: $confidence(X \Rightarrow Y, D) := \frac{support(X \cup Y, D)}{support(X, D)}$. The rule is called confident if $P(Y \mid X)$ exceeds a given minimal confidence threshold γ. The collection of frequent and confident association rules with respect to σ and γ is

$$R(D, \sigma, \gamma) := \{X \Rightarrow Y \mid X, Y \subseteq \Im, X \cap Y = \{\}, X \cup Y \in F(D, \sigma), confidence(X \Rightarrow Y, D) \geq \gamma\}.$$

Problem 2. (*Association Rule Mining*) Given a set of items $\Im$, a transaction database D over $\Im$, and minimal support and confidence thresholds σ and γ, find $R(D, \sigma, \gamma)$.

B. Previous works

The main pillar of ARM-algorithms is Apriori [4]. It is the best-known algorithm to mine association rules, which uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support. Over the years, a lot of improvements of Apriori, supported with

 ISBN: 978-1-61208-162-5

different types of memory structures, are proposed.

Recent ARM-algorithms, based on graph mining can be roughly classified into two categories. The first category of algorithms employs a breadth-first strategy. Representative algorithms in this category include AGM [5] and FSG [6]. AGM finds all frequent induced sub-graphs with a vertex-growth strategy. FSG, on the other hand, finds all frequent connected sub-graphs based on an edge-growth strategy. Algorithms in the second category use a depth-first search for finding candidate frequent sub-graphs. A typical algorithm in this category is gSpan [7], which was reported to outperform both AGM and FSG in terms of computation time.

A different approach for association rule searching is used in ECLAT [8]. It is the first algorithm that uses a vertical data (inverted) layout. The frequent itemsets are determined using sets of intersections in a depth-first graph.

In graph ARM approaches the bottleneck is the necessity of performing many graph isomorphism tests. To overcome this problem, alternative approaches use hash-based techniques for candidate generation. The representatives in this direction are DHP [9] based on direct hashing and pruning, [10] which proposed the use of perfect hashing, and IHP [11] that uses inverted hashing and pruning.

FP-Tree [12], Frequent Pattern Mining is another milestone in the development of association rule mining, which breaks the main bottlenecks of the Apriori. FP-tree is an extended prefix-tree structure storing quantitative information about frequent patterns. The tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of sharing nodes than less frequently occurring ones. The efficiency of FP-Tree algorithm has three reasons: (1) FP-Tree is a compressed representation of the original database; (2) it only scans the database twice; (3) it uses a divide and conquer method that considerably reduces the size of the subsequent conditional FP-Tree. The limitation of FP-Tree is its difficultness to be used in an interactive mining system, when a user wants to expand the dataset or change the threshold of support. Such changes lead to repetition of the whole mining process.

The Hmine algorithm [13] introduces the concept of hyperlinked data structure "Hyper structure" and uses it to dynamically adjust links in the mining process. Hyper structure is an array-based structure. Each node in a Hyper structure stores three pieces of information: an item, a pointer pointing to the next item in the same transaction and a pointer pointing to the same item in another transaction.

The innovation brought by TreeProjection [14] is the use of a lexicographical tree which requires substantially less memory than a hash tree. The number of nodes in its lexicographic tree is exactly that of the frequent itemsets. The support of the frequent itemsets is counted by projecting the transactions onto the nodes of this tree. This improves the performance of counting the number of transactions that have frequent itemsets. The lexicographical tree is traversed in a top-down fashion. The efficiency of TreeProjection can be explained by two main factors: (1) the transaction projection limits the support counting in a relatively small space; and (2) the lexicographical tree facilitates the management and counting of candidates and provides the flexibility of picking efficient strategy during the tree generation and transaction projection phrases.

Another data structure that is commonly used is a "trie" (or prefix-tree). Concerning speed, memory need and sensitivity of parameters, tries were proven to outperform hash-trees [15]. In a trie, every node stores the last item in the itemset it represents its support and its branches. The branches of a node can be implemented using several data structures such as hash table, binary search tree or vector.

Another algorithm for efficiently generating large frequent candidate sets, which use different data structures, is Matrix Algorithm [16]. The algorithm generates a matrix with entries 1 or 0 by passing over the cruel database only once, and then the frequent candidate sets are obtained from the resulting matrix. Finally association rules are mined from the frequent candidate sets. Experiment results confirm that the proposed algorithm is more effective than the Apriori.

This short overview of available algorithms and used structures shows the variety of decisions in association rule mining. As we can see graph structures, hash tables, different kind of trees, bit matrices, arrays, etc., are used for storing and retrieving the information.

Each kind of data structure brings some benefits and bad features. Such questions are discussed for instance in [17] where the comparison between tree-structures and arrays is made. Tree-based structures are capable of reducing traversal cost because duplicated transactions can be merged and different transactions can share the storage of their prefixes. But they incur high construction cost, especially when the dataset is sparse and large. Array-based structures incur little construction cost but they need much more traversal cost because the traversal cost of different transactions cannot be shared.

C. ArmSquare

Here, we offer one approach, which is focused on proposing appropriate coding of the items in database in order to use the possibilities of direct access to the information via coordinate vectors into multidimensional numbered information spaces. This structure combines the convenience of the work with array structures with economy and performance of tree structures, which lies in the ground of realized access method. The algorithm of obtaining association rules is very simple; we focus our attention over the possibilities of using such structures for storing information in data mining systems. In future more smart algorithms can be realized using multidimensional numbered information spaces as storage data structures.

The proposed approach is realized in the module ArmSquare as a part of Data Mining Environment PaGaNe [18]. ArmSquare is aimed to make analysis and monitoring over the produced association rules from frequent datasets. The main data structures in PaGaNe use the advantages of specific model for organization of the storage of information, called Multidimensional Numbered Information Spaces [19]. The model is realized in the access method ArM 32. Let us mention the existing confusion of abbreviation ARM used in literature for short denotation of "association rule miner" and

 ISBN: 978-1-61208-162-5

ArM, which means "Archive Manager". The name ArM was born in 1991 year (see [20]), two years before the defining of the association rule mining in [2]. The duplicating was used in the name of the realization: ArmSquare.

The rest of paper is organized in the following way. In Section 2, a brief description of Multidimensional Numbered Information Spaces is given. Sections 3 and 4 present our approach and program realization, which are based on the given possibilities for direct access to the points of multidimensional numbered information spaces. The differences between proposed algorithm and existing ones are discussed in Section 5. A short explanation of used databases from different fields is given in Section 6. Finally, some conclusions are highlighted.

II. MULTIDIMENSIONAL NUMBERED INFORMATION SPACES

Following the Multi-Domain Information Model (MDIM), presented in [19] and realized by ArM 32, the elements are organized in a hierarchy of numbered information spaces with variable ranges, called ArM-spaces.

A. Constructs

There exist two main constructs in MDIM – *basic information elements* and *numbered information spaces*. Basic information element is an arbitrary long sequence of machine codes (bytes). Basic information elements are united in numbered sets, called numbered information spaces of range 1. The numbered information space of range n is a set, which elements are numerically ordered information spaces of range $n-1$. ArM32 allows using of information spaces with different ranges in the same file.

Every element may be accessed by correspond multidimensional space address (ArM-address) given by a coordinate array. The coordinate array is represented as numerical vector $A=(n,p_1,...,p_n)$, in which starting position shows the dimension of the space and next positions contains the coordinates of the points, thorough which the information can be reached. Sometimes, accounting the difference between the meaning of starting coordinate and next coordinates, the vector is written as $(n:p_1,...,p_n)$.

Another constructs, connected with MDIM are *indexes* and *metaindexes*. Every sequence of space addresses $A_1,A_2,...,A_k$, where k is an arbitrary natural number, is said to be an *index*. Every index may be considered as basic information element and may be stored in a point of any information domain. In such a case, it will have a space address which may be pointed again. Every index which point only to indexes is said to be a *meta-index*.

Special kind of space index became the *projection*, which is analytical given index. There are two types of projections: (1) *Hierarchical projection* – in which the top part of coordinates is fixed and low part vary for all possible values of coordinates, where non-empty elements exist; and (2) *Arbitrary projection* – in this case is possible to fix coordinates in arbitrary positions and the rest coordinates vary for all possible values of coordinates, where non-empty elements exist.

B. Operations

It is clear that the operations are closely connected to the defined structures. So, we have operations with:

- *basic information elements*: Because of the rule for existing of the all structures given above we have need of only two operations: updating and getting the value and two service operations: getting length and positioning in the element;
- *spaces*: With two spaces we may provide two operations: copying the first space in the second and moving the first space in the second with modifications specifying clearing or remaining the second space before operation;
- *indexes and meta-indexes*: Using the hierarchical projection we may crawl the defined area and extract next or previous empty or non-empty elements as well as to receive the whole index or its length, of the non-empty elements, which addresses fall into defined projection. The same operations (but only for non-empty elements) can be made for arbitrary projection. The operations between indexes are based on usual logical operations between sets. The difference from usual sets is that the information spaces are built by interconnection between two main sets: set of co-ordinates and set of information elements.

III. ALGORITHM DESCRIPTION

We propose to use the abilities of multidimensional numbered information spaces for storage the information about interconnections between items and their combinations for facilitating association rule mining.

The proposed algorithm makes special analysis for each transaction and stores the frequency information in ArM-space, using the possibility of these spaces for accessing to the data via coordinate arrays. The algorithm consists of three phases: (1) pretreatment; (2) data processing and (3) analysis and monitoring.

A. Pretreatment

In the pretreatment phase, the following steps are made:

- The transactions of the incoming dataset D may be split into subsets, $D_b, b=1,...,d \quad \bigcup_{b=1}^{d} D_b = D$ by some condition (periods, regions, etc.). The mapping between the names of these groups and natural numbers $b=1,...,d$ is made;
- Creating a mapping between incoming items $i_j \in \Im$ and natural numbers $c_j \in \mathbb{N}$ by order of first occurrence of the item. This way if $\Im$ is a set of items, then $\overline{\Im}$ is also a set of items that contains the numbers from 1 to n ($n=|\Im|$), which code the items of $\Im$. Each incoming transaction $T=(tid,I), \ I=\{i_1,...,i_k\} \subseteq \Im$ is transformed into $\overline{T}=(tid,C), \quad C=\{c_1,...,c_k\} \subseteq \overline{\Im}$. The transaction database D over $\Im$ is transformed to $\overline{D}$ over $\overline{\Im}$;
- The items in each transaction $\overline{T}$ are sorted in increasing order. The received transaction is denoted by $\vec{T}$. Ordering the items in the transaction has a great importance for the consequent steps.

B. Data processing

The intermediate phase is data processing. The data processing is closely depended on the length of the derived association rules. The greater the length, the more resources are needed. Because of this usually a special parameter $MaxK$ is used for limiting the maximum length of examined association rules. The algorithm traverse all combinations from 1 to $MaxK$. Let k be the examined number of items $1 \le k \le MaxK$. For each transaction $\vec{T}$, $n = |\vec{T}| \ge k$ we make all possible combinations $Z^l = \{c_1^l, ..., c_k^l\}$, $Z \subseteq \vec{T}$, $l = 1, ..., \frac{n!}{k!(n-k)!}$. The element of Arm-spaces with co-ordinates $(c_1^l, ..., c_k^l)$ accumulates the number of occurrence of corresponded itemset $Z^l = \{c_1^l, ..., c_k^l\}$ (Fig. 1).

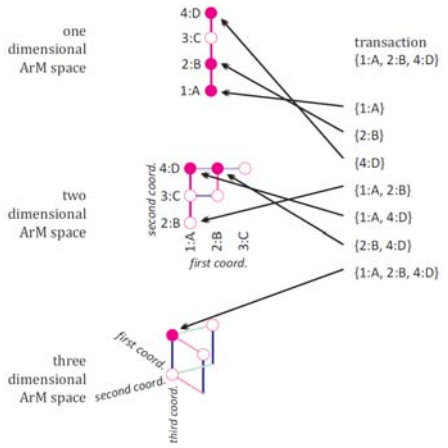


Figure 1. Accumulating in ArM spaces of the number of occurrence of produced itemsets from one transaction

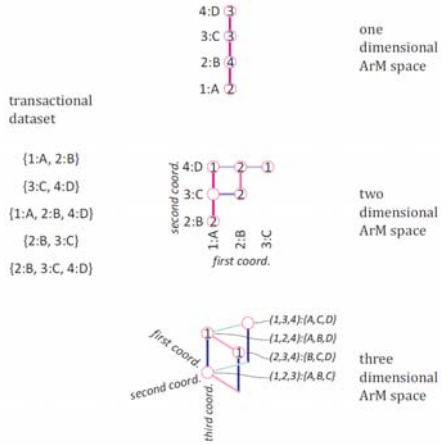


Fig. 2. Result of data processing of the database

In the case when D is split in subsets $D_b, b = 1, ..., d$ additional coordinate in the space address is placed for marking the number of the group b where transaction belongs to and the space address became following form $(b, c_1^l, ..., c_k^l)$.

As far as the processing over combinations with different lengths as well as subsets of database D are independent, operations may be provided in parallel.

Finally, the support of the itemsets, which are driven from the transactions, is accumulated in the corresponded points in ArM-spaces (Fig. 2). Note that because of the ordering, not all coordinates in corresponded space are used. ArM 32 does not waste memory for empty points.

C. Analysis and monitoring

The analysis is made over the itemsets with particular length k, $1 \le k \le MaxK$ and using:

- A minimal support σ, which the itemsets to be included in the resulting list must have;
- A minimal confidence γ, which the association rules, created on the basis of the already selected itemsets, must have.

For obtaining all existing k-itemsets, whose support are at least σ, a traversal of all non-empty elements in a k-dimensional ArM-space is made. The coordinates of each element $(c_1^l, ..., c_k^l)$, which contains value, no less than σ, corresponds to itemset $Z^l = \{c_1^l, ..., c_k^l\}$, which is included in the resulting list $F(D, \sigma)$.

In the case where a database is split into groups, the traversal is made for each group taking into account that the first coordinate indicates the number of the group. The support of itemset $Z^l = \{c_1^l, ..., c_k^l\}$ for the whole database is received as a sum of values, contained in the points with corresponding coordinates in each group $(b, c_1^l, ..., c_k^l), b = 1, ..., d$.

One itemset $Z^l = \{c_1^l, ..., c_k^l\}$ is a source of producing several association rules.

Let $Z = \{c_1, ..., c_k\}$ be a k-itemset, $Z \in F(D, \sigma)$. The collection of association rules

$$R = \{X \Rightarrow Y \mid X, Y \subseteq \overline{\Im}, X \cap Y = \{\},$$
$$X \cup Y = Z, confidence(X \Rightarrow Y, D) \ge \gamma\}$$

is obtained by examining all possible combinations with length from 1 to k-1, which is given as a body of the rule $X^j = \{c_1^j, ..., c_p^j\}$, $p = 1, ..., k-1, X^j \subset Z$. The rest of the items forms the head of the rule $Y^j = Z \setminus X^j$. For an association rule $X^j \Rightarrow Y^j$:

- from the point with the space address $(c_1^j, ..., c_p^j)$, which correspond to X^j, the $support(X^j)$ is received. Taking into account that the body is part of an already existing itemset, this value is more than zero. The confidence of this association rule is calculated as $confidence(X^j \Rightarrow Y^j) = \frac{support(Z)}{support(X^j)}$;
- if $confidence(X^j \Rightarrow Y^j) \ge \gamma$, then $X^j \Rightarrow Y^j$ is included in the list of resulting association rules $R(D, \sigma, \gamma)$.

 ISBN: 978-1-61208-162-5

IV. PROGRAM REALIZATION

The proposed algorithm was realized as analyzing tool in data mining environment PaGaNe.

A. Input data

The system allows creation of a new database as well as adding new transactions to the same or different groups in an already existing database. During the input, the system accumulates the information for maximal length and average length of the transactions by each group separately. The repeated elements within the transaction are omitted.

Each transaction in the system is presented as numerical vector, with the length equal to the number of the elements, which participate in the transaction. The elements in the vectors are numbers, which correspond to the position of the element in dynamically expanded nomenclature, which contains the names of the items. Finally, these numbers are sorted. Sorting the elements in the vectors has a great importance for the consequent steps.

B. Pretreatment in ArmSquare

Before starting the processing, one has to give a maximal number of combinations, which will be interesting – *MaxK*. Usually not all combinations between elements are interesting, but only a limited number of them – 2, 3, or 4, and no more than 10. The pretreatment performs a special analysis for each transaction in all groups and stores the frequency information in ArM-spaces, using the ability of the ArM-spaces for accessing the data via coordinate arrays. The user is interested in combinations from 1 to *MaxK*.

Let us trace the process for the transaction $\vec{T}$, which belongs to the group with number b. The length $n = |\vec{T}|$ varies depending on the numbers of the elements that were in the transaction. The system loops by k from 1 to $\min(n, MaxK)$ in order to traverse all possible combinations. Each combination $\{c_1,...,c_k\}$ is used to form ArM-address for "k+1" dimensional space $(b, c_1,...,c_k)$ and at this point a support value is incremented by 1.

C. Analysis and monitoring

The user can choose which length of itemsets he wants to observe. This length can vary between 2 and *MaxK*. Other parameters are minimal support and minimal confidence for a given database.

For obtaining all existing k-itemset with a support of at least σ, crawling over the k-dimensional ArM-space is done using the function *ArmNextProj*, starting with hierarchical projection $(-,-,...,-)$. In case of observing only a concrete group b, the crawling is made in k+1-dimensional ArM space with starting projection $(b,-,...,-)$. Using the function *ArmRead* for the current extracted non-empty element, the value is read and is compared with σ. If this value is no less than σ, the corresponded itemset is included into the resulting list of itemsets $F(D,\sigma)$. The resulting itemsets $F(D,\sigma)$ is sorted by decreasing support.

For receiving all association rules, created on the basis of itemsets from $F(D,\sigma)$, which have a confidence no less than γ, for each itemset $Z = \{c_1,...,c_k\}$ where $Z \in F(D,\sigma)$, every possible combination with a length from 1 to k-1, where $X^j = \{c_1^j,...,c_p^j\}$, $p = 1,...,k-1, X^j \subset Z$ is assumed as a body, while the rest of the items are taken as a head of the rule $Y^j = Z \setminus X^j$, is examined. For association rule $X^j \Rightarrow Y^j$:

- the value of $support(X^j)$ is received using function *ArmRead* from coordinate space address $(c_1^j,...,c_p^j)$, which correspond to the body X^j;
- the $confidence(X^j \Rightarrow Y^j) = \dfrac{support(Z)}{support(X^j)}$ is calculated;
- the association rule $X^j \Rightarrow Y^j$, whose confidence is no less than γ is included in the list of resulting association rules $R(D,\sigma,\gamma)$.

Optionally, association rules can be sorted by decreasing confidence.

Using the ArM functions, the following additional operations, which can be used for analysis of the database, can be executed:

(1) Observation of all itemsets with given length k. For this purpose a function *ArmProjIndex* with hierarchical projection on the highest level in k-dimensional space is used. In the case of viewing the itemsets, belonging to a given group – the projection is one level lower in k+1 dimensional space with the highest coordinate equal to the number of the group fixed.

(2) Looking for k-itemsets, containing concrete item with given number c. A loop from 1 to k allows crawling the arbitrary projection (-,…,-,c,-,…,-), where position of c varies accordingly to the loop phase. A function *ArmProjIndex* uses these projections and extracts all non-empty elements, which define the corresponding itemset. The union of all these elements is the result of the request.

D. Advanced specifics of ArmSquare

In Apriori algorithm min-support is set globally for combinations with different lengths. In our algorithm, after building the spaces, statistics for min-support for each area can be derived separately (the amount of space is equal to the number of elements in combinations), which allows to give for further analysis different min-support for different numbers of elements in combinations.

In a higher value of min-support Apriori is highly convergent and reaches a relatively short itemsets, where a small amount of min-support is close to total exhaustion of short itemsets.

Structuring the support of the itemsets in ArM-space allows subsequent analysis to be made very quickly by setting a different min-support and profiles of different lengths of itemsets, while other ARM-approaches derive all successive combinations in ascending order and changing the min-support causes a repetition of the whole algorithm.

 ISBN: 978-1-61208-162-5

The information for itemsets with particular length containing a specific element can be directly extracted.

The database can be interactively expanded as well as the processing of the transactions can be made in parallel.

V. IMPLEMENTATIONS

The realized tool allows different types of useful implementations in a wide spectrum of applications.

In one experiment we have used a retail market basket data set supplied by anonymous Bulgarian retail supermarket store. The data was collected over one year period (2008 year) from purchasing in a middle supermarket in a town with about 30 000 citizens. The total number of transactions was 108 846. The number of items were 3 609. The maximum length of the transactions was 23. The average items into transactions were 2.76 items per transaction. The transactions were grouped in accordance of months, when corresponded purchase was made. Several experiments were conducted with this dataset. Using all transactions, the most frequent combinations with 2, 3 and 4 length were extracted. Also there was used the possibilities of the ArmSquare to analyze different bins one to others and to extract the deviations of purchasing during the months.

Other experiments were made over a dataset that included several types of color harmonies and contrast features, extracted by 600 paintings of 19 artists from different movements of West-European fine arts and Eastern Medieval Culture. The pictures were obtained from different web-museums sources using ArtCyclopedia as a gate to the museum-quality fine art on the Internet. Each row of formed dataset contained the name of the artists, followed with harmonies and contrast features, presented in the manner of transactional dataset: "feature name"="value". Using the possibility of binning the dataset by class label allowed to use ArmSquare as element in the generation rule phase of CAR-algorithm and extract typical combinations of features for examined artists [21]. The constructed classifier, based on ArmSquare, outperforms classifiers with similar classification models (OneR, J48, JRip), realized in Weka.

VI. CONCLUSIONS AND FUTURE WORKS

The main focus in the realization of ArmSquare is to show the possibilities to use the advantages of multi-dimensional information spaces for memory structuring in the area of data mining and knowledge discovery. The variety of the tasks that can be made with proposed frequent association rule miner ArmSquare allows comprehensive and facile analysis of the situations and conducting forecasting in wide areas of applications. The next steps will be focused on improving the algorithm of extracting rules, especially realizing the ARUBAS algorithm [22] over the multi-dimensional information spaces.

ACKNOWLEDGMENT

This work was supported in part by Hasselt University under the Project R-1876 and by Bulgarian NSF under the project D002-308.

REFERENCES

[1] Bodon, F., "A fast APRIORI implementation". In IEEE ICDM Workshop on FIMI, Melbourne, Florida, USA, 2003.

[2] Agrawal, R., Imieliński, T., and Swami, A., "Mining association rules between sets of items in large databases". In Proc. of the ACM SIGMOD ICMD, Washington, DC, 1993, pp. 207-216.

[3] Goethals, B., Efficient Frequent Pattern Mining. PhD thesis in Transnationale Universiteit Limburg, 2002.

[4] Agrawal, R. and Srikant, R., "Fast algorithms for mining association rules", In Proc. of the 20th Int. Conf. on VLDB, 1994, pp. 487-499.

[5] Inokuchi, A., Washio, T., and Motoda, H., "Complete mining of frequent patterns from graphs: mining graph data". In Machine Learning, Vol.50, 2003, pp. 321-354.

[6] Kuramochi, M. and Karypis, G., "Frequent subgraph discovery". In Proc. of the 1st IEEE Int. Conf. on DM, 2001, pp. 313-320.

[7] Yan, X. and Han, J., "gSpan: Graph-based structure pattern mining". In Proc. of the 2nd IEEE Int. Conf. on DM, 2002, pp. 721-724.

[8] Zaki, M., Parthasarathy, S., Ogihara, M., and Li, W., "New algorithms for fast discovery of assoctiation rules". In Proc. of the 3rd Int. Conf. on KD and DM, 1997, pp. 283-286.

[9] Park, J., Chen, M., and Yu, P., "An effective hash based algorithm for mining association rules". In Proc. of ACM SIGMOD Int. Conf. on Management of Data, 24/2, 1995, pp. 175-186.

[10] Özel, S. and Güvenir, H., "An algorithm for mining association rules using perfect hashing and database pruning". In Proc. of the TAINN, 2001, pp. 257-264.

[11] Holt, J. and Chung, S., "Mining association rules using inverted hashing and pruning". Information Processing Letters Archive, 83/4, 2002, pp. 211-220.

[12] Han, J. and Pei, J., "Mining frequent patterns by pattern-growth: methodology and implications. In ACM SIGKDD Explorations Newsletter 2/2, 2000, pp. 14-20.

[13] Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., and Yang, D., "Hmine: hyper-structure mining of frequent patterns in large databases". In Proc. of IEEE ICDM, 2001, pp. 441-448.

[14] Agarwal, R., Aggarwal, C., and Prasad V., "A tree projection algorithm for generation of frequent item-sets". In Journal of Parallel and Distributed Computing, 61/3, 2000, pp. 350-371.

[15] Bodon, F. and Ronyai, L., "Trie: an alternative data structure for data mining algorithms". In Mathematical and Computer Modelling, 38/7, 2003, pp. 739-751.

[16] Yuan, Y. and Huang, T., "A Matrix algorithm for mining association rules". In LNCS, Vol. 3644, 2005, pp. 370-379.

[17] Liu, G., Lu, H., Yu, J., Wang, W., and Xiao, X., "AFOPT: An efficient implementation of pattern growth approach". In Workshop on Frequent Itemset Mining Implementation. (FIMI 03), 2003.

[18] Mitov, I., Ivanova, K., Markov, K., Velychko, V., Vanhoof, K., and Stanchev, P., "PaGaNe – a classification machine learning system based on the multidimensional numbered information spaces". In WSPS on CEIS, No. 2, 2009, pp. 279-286.

[19] Markov, K., "Multi-domain information model". In Int. J. on Information Theories and Applications, 11/4, 2004, pp. 303-308.

[20] Markov K., Ivanova, K., Mitov, I., and Karastanev, S., "Advance of the access methods". Int. J. on Information Technologies and Knowledge, 2/2, 2008, pp. 123-135.

[21] Ivanova K., Stanchev P., and Vanhoof K., "Automatic tagging of art images with color harmonies and contrasts characteristics in art image collections". Int. J. on Advances in Software, 3/3&4, 2010, pp. 474-484.

[22] Depaire, B., Vanhoof, K., and Wets, G., "ARUBAS: an association rule based similarity framework for associative classifiers". In IEEE Int. Conf. on Data Mining Workshops, 2008, pp. 692-699.

 ISBN: 978-1-61208-162-5

Review of Shape-based Similarity Algorithms and Design Retrieval Methods for Computer-aided Design and Manufacturing

Leila Zehtaban and Dieter Roller
Institute of Computer Aided Product Development Systems
Universität Stuttgart, Stuttgart, Germany
Email: zehtaban@informatik.uni-stuttgart.de, roller@informatik.uni-stuttgart.de

Abstract— **Reusing engineering data has opened a new opportunity to improve product quality, shorten design lead-time and reduce costs using existing know-how within the design process. Geometrical aspects or 3D shape information of a product is an essential data which can be reused in CAD software. In order to compare and retrieve the existing 3D models, having a precise computational representation of a shape, so-called shape index or shape signature, is a main challenge. The shape signature is often used for the shape similarity comparison. There are several specifications for a shape signature like quick to compute, easy to index, invariant under transformation, independent of 3D representations, tessellation, genus or topology. The algorithms or the methods which decompose a shape into a signature can be classified into seven main classes. This paper aims to focus on the discussion of the first three methods, i.e., Invariant-based methods, Harmonics-based methods, and Graph-based methods, and provide the related literature review on their underlying approaches with highlighting methodologies, advantages and disadvantages.**

Keywords-shape signature; shape similarity comparison; 3D shape retrieval; reused design.

I. INTRODUCTION

New technological progress has enabled Computer-aided Design (CAD) software to incorporate engineering know-how into the design process in order to improve product quality, shorten design lead-time and reduce costs. Any manufacturer has an accumulated amount of know-how related to design, production and performance of existing or previously manufactured products. Accessibility and the possibility of reusing this accumulated knowledge is a key factor ***for*** optimizing design and performance of a new product. Using a capable procedure for identifying similarity between a new possible product with items listed in the existing product data-bases enables a design engineer to find a professional base to design a new product. The new design can be well optimized using existing know-how of the existing product in design, production and performance.

The similarity comparison between two objects could include diverse similarity aspects like similarity in shape (structure), design intent (functionality), production specifications, etc. However, the shape similarity comparison is one of the most important bases for any comprehensive similarity comparison in product design. The complexity of shape similarity comparison arises from the challenge of finding a computational representation (signature) for a shape which can be applied for the shape similarity comparison. The current shape similarity methods can be classified as follows [1]: Invariant-based methods, Harmonics-based methods, Graph-based methods, Statistics/probability-based methods, 3D object recognition-based methods, Feature recognition-based methods and Group Technology-based methods.

In this paper, we highlight the first three classes by having a literature review on their different underlying methods. Although the methods which are classified under the same classification, originally apply an identical concept to decompose a shape into a signature, nevertheless there are still differences regarding the utilized techniques. In the following, Section II describes the invariant-based methods, Section III describes the harmonics-based methods, Section IV describes the graph-based methods, and finally, Section V summaries the paper.

II. FIRST CLASSIFICATION: INVARIANT-BASED METHODS

These approaches use invariants or descriptors of the 3D shape such as volume, surface area, aspect ratio, higher order moments or moment invariants as signatures [1]. At the following four methods which belong to the category of invariant-based will be briefly discussed. These methods include: RTS-invariants, Moments and relevance feedback, Non-dimensional and scale-independent features, and Elementary-shape-based features and active learning will be explained.

A. *RTS-invariants*

In the method from Cybenko et al. [2], solid objects given in a standard digital representation like the IGES file format are converted into a surface triangular mesh representation. Afterwards, the triangular mesh representation is converted into a voxel model representation using a flood filling method. For the shapes represented by voxel model, geometrical moments are calculated and used to normalize the object into a canonical form. Shape features are computed by calculating variant volumetric invariants. They are called RTS-invariants because these features are invariant against rotation, translation and scaling. The following RTS-invariants can be calculated: second-order

 ISBN: 978-1-61208-162-5

3D moments invariants, spherical-kernel moments invariants, axis aligned bounding box and centroid of the object, and the surface area of the objects. In the first step of similarity measurement, feature vectors are used to compute a set of best candidate objects. On this set of the best candidates a voxel-by-voxel comparison is performed as the second step of similarity measurement. This step allows a detailed comparison between voxel model representation of objects and is based on template matching.

B. Moments and relevance feedback

Elad et al. [4][5], used moments as shape features of 3D models and relevance feedback as an iterative and interactive method to improve the performance retrieval. For the models given in VRML file format (Virtual Reality Modeling Language), the geometrical moments are calculated and approximated up to the third-order and used to normalize the object in a canonical form. For the normalized objects moments are approximated again (up to forth-to-seventh-order is sufficient) and used as a feature vector of objects.

In the first step of similarity measurement a set of the best candidates is presented to the user by computing the Euclidean distance between feature vectors of the primary object and objects from the database. After that the user has the ability to influence the future search results by applying the method of relevance feedback. The user can mark a subset of presented results as relevant or as irrelevant. Based on these markings, which capture the user-perceived similarity between objects, the distance measure can be adapted and a new search results calculated. The adaption of the distance measure is based on Support Vector Machine (SVM) learning algorithms and can be repeated until the user is satisfied with the search results.

C. Non-dimensional and scale-independent features

Rea et al. [6] and Corney et al. [7], used various non-dimensional and scale-independent features as signature for 3D CAD models in an internet search engine. Most of these features are computed using object characteristics such as volume, surface area and convex hull of objects.

For example, the features like crinkliness, compacness, hull crumbliness, etc. are calculated as following [7]: Crinkliness is defined as the surface area of the model divided by the surface area of the sphere having the same volume as model. Compactness is defined as a ratio of the volume squared over the cube of the surface area and used as a non-dimensional feature. Hull crumbliness is defined as a ratio of objects surface area to the surface area of its convex hull. Hull packaging is defined as the percent of the convex hull volume not occupied by the original object. Hull compactness is defined as ratio of the convex hull's surface area cubed over the volume of the convex hull squared.

Further features being used are: ratio of the longest edge to the shortest edge of the bounding box, number of the holes of the object and number of the facets of the object. User can specify combination of features which are used in the similarity search and tolerance values for these features.

D. Elementary-shape-based features and active learning

The method from Zhang et al. [8][9], describes that features such as volume, surface area, moments and Fourier transform coefficients can be well extracted from a mesh representation and be considered as the signature of an object. The inspiration of this method is to compute features for elementary shapes such as triangles and tetrahedrons in advance and sum up the feature values of the elementary shapes in order to get the feature value of the whole object. Annotation of the object was used as a method to improve the performance retrieval. The hidden annotation has to be performed as a learning stage before a database can be used for the similarity search. By using an active learning method the system determines the sample objects to the annotator. The sample objects are selected so that annotation of the object can provide the maximum information or knowledge gain to the system.

Using this method reduces the number of training samples by selecting the most informative ones to the annotator.

E. Evaluation of invariant-based methods

All invariant-based methods have the advantage of being robust to small changes in shape. The disadvantage of these methods refers to the improbable partial matching.

The method from Cybenko et al., suffers from the requirement of a huge storage requirement for every object and its different models. In addition, the voxelization of models is a time and memory consuming process. [2]

In the method from Elad et al., with using the relevance feedback, not only the geometrical similarity is being computed between the objects, but also the user-perceived similarity can be incorporated in the similarity search process. Hence, the retrieval performance is being improved by retrieving more objects than user has in mind. [4][5]

The method from Corney et al. and Rea et al., can be useful as a coarse filter in huge databases. However, to perform a finer comparison between objects when the sets of retrieved objects are large, combination of this method with further methods might be necessary. [6][7]

The disadvantage of the method from Zhang et al., lies on the requirement of an explicit routine to compute a feature value for elementary shape. Nevertheless, it is difficult to develop explicit routine to compute the high order moments for triangles and tetrahedrons. Zhang et al., used the method of hidden annotation and active learning to improve the retrieval performance of the system. In practice, an annotation of large databases can hardly be performed because of the manual effort. Besides, with applying partial annotation, it is difficult to decide how much annotation is sufficient for specific database. [8][9]

III. SECOND CLASSIFICATION: HARMONIC-BASED METHODS

These methods use a set of harmonic functions of a shape as signature. Spherical or Fourier functions are usually used to decompose a discrete 3D model into an approximate sum of its (first n) harmonis components [1]. The four methods of this category will be discussed in the following sections:

 ISBN: 978-1-61208-162-5

Ray-based SH-descriptor, Rotation-invariant SH-descriptor, Layered depth sphere-based SH-descriptor; and Concrete radialized spherical projection descriptor.

A. *Ray-based SH-descriptor (Spherical Harmonics)*

In the method from Vranic [10][11], 3D models represented by polygon meshes are normalized to achieve invariance against rotation, translation and scaling. For that purpose a Principal Component Analysis (PCA) which can be applied to a discrete set of points, as well as the union of all polygons of the mesh with infinitely number of points. After normalization, the 3D models are characterized by defining a function on a sphere which measures the extension of an object in different directions.

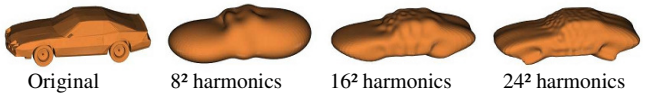


Figure 1. Multi-resolution representation used to derive feature vectors from Fourier coefficients for spherical harmonics [10]

For each direction a ray is casted from the center of mass in order to compute the last point of the intersection with the polygonal mesh which is used as a sample of the function. After sampling the function Fast Fourier Transformation (FFT) is performed to obtain the Fourier coefficients to be applied as feature vector. Figure 1 represents reconstruction of the different levels of a primary object when using three different spherical harmonics coefficients.

B. *Rotation-invariant SH-descriptor*

Kazhdan et al. [3] claim that the methods using PCA are unstable referring to the multiplicity of eignevalues and its sensitivity to outliers.

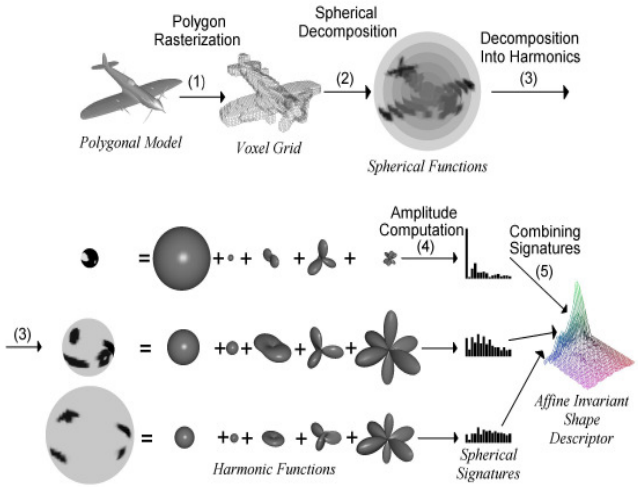


Figure 2. Computing the Harmonic Shape Representation [3]

As a solution a new method [12][13], to compute the harmonic shape representation is proposed. In this method, the model polygon is rasterized into a $64 \times 64 \times 64$ voxel grid. The voxel grid is decomposed into 32 functions on concentric spheres by restricting the voxel grid to spheres with radii 1 to 32. By decomposing each of these functions as a sum of its first 16 harmonic components, analogous to a Fourier decomposition into different frequencies and define the signature of each spherical function as a list of these 16 norms and combining the different signatures, a 32×16 signature for 3D model is obtained. In order to compare two harmonic presentations, the Euclidean distance between them should be computed. An example of the explained method is shown in Figure 2.

C. *Layered Depth Spheres (LDS)-based SH-descriptor*

Vranic [14][15], described a further harmonics-based method which captures information about internal structure of objects. The shape descriptor is extracted from a triangle mesh representation of the objects. Invariance against translation and scaling is achieved using Continues PCA (CPCA). 3D model is decomposed into a family of function on the sphere restricting function values by lower and upper bounds which describe a bounded area of the model. Using ray cast method for rays emanating from the origin in many directions all points of intersection with the polygonal mesh are computed.

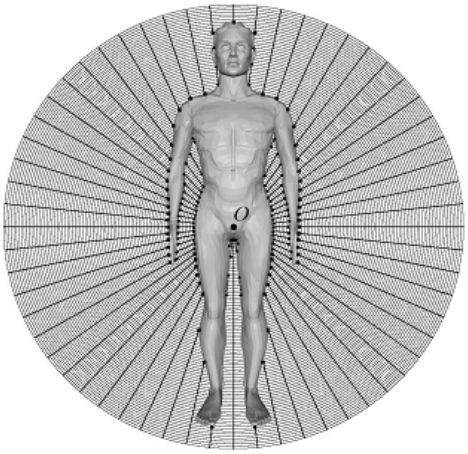


Figure 3. Concept of Layered depth Sphere with an example in 2D [15]

For intersection points the closest sphere and a set of corresponding value of the function on that sphere is determined. If two intersection points of the same ray belong to the same sphere then the larger distance determines the function value. On each sampled function on the sphere Fast Fourier Transformation (FFT) is performed to obtain a set of coefficients. The PCA method can be performed during the normalization step or the properties of spherical harmonics can be used to achieve rotation invariance.

D. *Concrete Radialized Spherical Projection Descriptor (CRSP)*

In the method from Papadakis et al. [16], a shape descriptor is extracted from a triangle mesh representation of 3D models. In this method, scaling and axial flipping invariance is achieved referring the properties of spherical harmonics.

 ISBN: 978-1-61208-162-5

Rotation and translation invariance is achieved by applying CPCA and Principal Component Analysis on the model's Normal (NPCA). This algorithm, results in two versions of an object and, therefore, two descriptors for an object. For each version of the object a set of functions on spheres is defined, which are sampled by casting rays from the origin of the object.

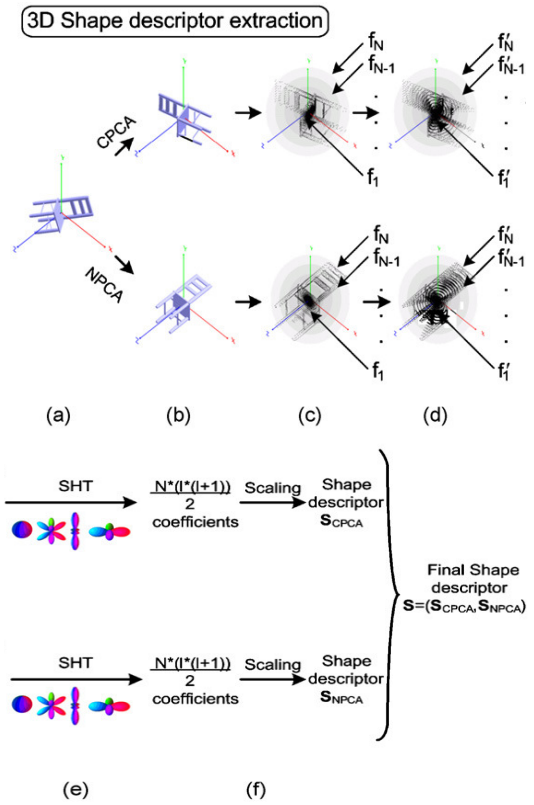


Figure 4. The stage of the shape matching using CPCA and NPCA [16]

A function on a sphere represents intersection points of the models surface with rays and also all points in the direction of each ray that are closer to the origin than the furthest intersection point. For every function Short-time Fourier Transform (SFT) is performed to obtain the Fourier coefficients. Scaling invariance of the descriptor is achieved using properties of spherical harmonics. Figure 4 illustrates the stages of the shape decomposition and matching as well as obtaining the shape descriptor/signature respectively.

E. Evaluation of Harmonics-based methods

All harmonics-based methods have an advantage which feature extraction and similarity measurements are efficiently performed. Drawbacks of these methods are as following: first; specific details of shape can not be captured, and second; partial matching is not possible in these methods. [1]

Kazhdan used a coarse voxel grid to achieve robustness against small changes of shape. However, coarse voxel grid causes loss of many details. [11] Voxelization also affects efficiency of feature extraction.

The ray-based method allows an embedded multi-resolution representation of the descriptor. This means that a descriptor contains all descriptors having lower dimension. [11]

Unlike the ray-based method, LDS-based method captures information about the internal structure of objects by defining several functions on spheres instead of only one.

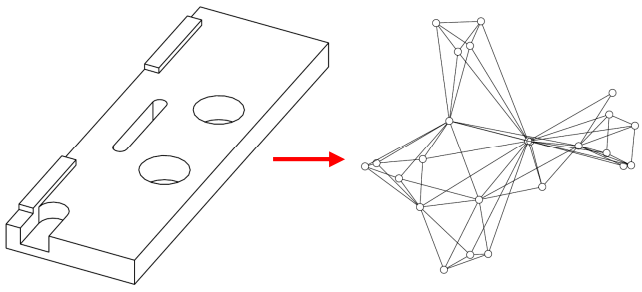
The CRPS method improves the invariance properties of the descriptor by applying two normalization methods, CPCA and NPCA. Thus, the retrieval performance of the descriptor is improved. Although this process increases the complexity of descriptor, since for each object two descriptors are extracted. [16]

IV. THIRD CLASSIFICATION: GRAPH-BASED METHODS

In Graph-based approaches sub-graph isomorphism is used in order to match B-Rep graphs, or to match eigenvalues of a model signature graph which is constructed from the B-Rep graph. Five different methods belonging to the graph category are briefly explained in the following [1]. These five methods include: Model signature graphs, Attributed graphs, Reeb graphs and Skeletal graph with parameter-controlled thinning.

A. Model signature graphs

McWherter et al. [17][18][19][20], developed Model Signature Graph (MSG) for similarity measurement between 3D CAD models. MSGs are labeled, undirected graphs, which are generated using the Boundary Representation (B-Rep) on CAD models. The boundary representation consists of a set of edges and a set of faces. For the definition of MSGs every face of the model is represented as a graph vertex and every edge in the B-Rep is represented as a graph edge. Labels of edges and vertices contain attributes of faces and edges in B-Rep, such as topological identifier, underlying geometrical representation, etc. For every MSG the eigenvalue spectrum and Invariant Topology Vector (ITV) are extracted and used to perform similarity comparison. ITV contains graph invariants, such as vertex and edge counts, maximum, minimum, mode degrees graph diameter, etc.



(a) Solid Model (b) Model Signature Graph

Figure 5. A model and its transformation into a Model Signature Graph [19]

 ISBN: 978-1-61208-162-5

The spectrum of a graph is sorted as the eigenvalues of its adjacency matrix, which holds information related to the graph structure. In addition, the eigenvalues of the graphs can be used to partition the graph into two or more sub-graphs in order to compare substructure of the graphs.

B. Attributed graphs

In the approach of El-Mehalawi und Miller [21][22], attributed graphs are used as signature of 3D CAD models, which are extracted from STEP files of these models. Attributed graphs are quite similar to MSGs, hence, graph nodes describe the faces of CAD components and graph edges describe the edges of CAD components. In addition, the node attribute correspond to the surface attribute, such as type of surface and direction of the normal. Edge attributes correspond to the edge attributes in B-Rep, such as type of the edge, direction of the normal and length of the edge.

For the retrieval process, abstract information is extracted from attributed graphs and used in the first step of the retrieval process. The abstract data is the total number of nodes, number of nodes representing plan, cylindrical and conical surfaces. These data are used as an index, where the set of graphs candidate similar to the query graph can be calculated very quickly. In addition, a more accurate comparison is applied for the set of candidate graphs. To finalize the method and to complete the retrieval procedure, inexact graph matching algorithm based on an integer programming model is applied. This algorithm has a polynomial computational complexity.

C. Skeletal graphs with thinning

Sundar et al. [23][24], used skeletal graphs as signature of 3D models for similarity measurement as figure 6 present it.

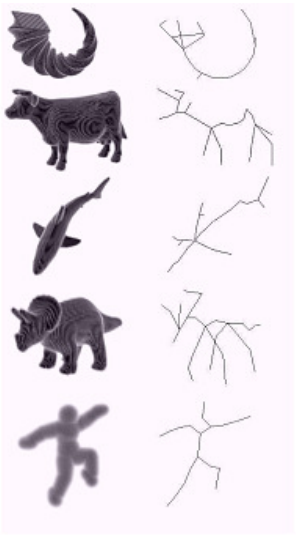


Figure 6. Skeletal graphs based on the thinning algorithm [23]

To extract the skeletal graph of an object, the belonging 3D model ought to be converted into a voxel model. In the next step, for the skeletonization process, a parameter-controlled thinning algorithm is used to calculate a subset of voxels. In this thinning method the thickness of the skeleton is determined by the parameter given the user. Hence, a family of the different voxel sets can be calculated, each one is thinner than its parent. The thinness parameter classifies the importance of the voxels for the boundary coverage by comparing the distance transform of the voxel with its 26 neighbors. After skeletonization the Minimum Spanning Tree (MSN) algorithm is applied in order to generate an undirected acyclic graph out of unconnected skeletal points. For every node in the graph Topological Signature Vector (TSV) is defined which holds information related to the node underlying sub-graphs structure.

TSV contains the eigenvalues of the sub-graph's adjacency matrix and is used as an index to fast determination of a set of best candidate graphs. On the set of candidate graphs a graph matching algorithm is performed by reformulating the problem of largest isomorphic sub-graph as the problem of finding the "maximum cardinality, minimum weight matching" in a bipartite graph. To preserve the hierarchical structure of the graph a greedy form of the above bipartite formulation is combined with a recursive depth search.

D. Skeletal graphs with parameter-based thinning

In the method from Iyer et al. [25][26][27], skeletal graphs and feature vectors jointly present the signature of 3D models. 3D models are normalized into a canonical form and converted into voxel model. In the skeletonization process, iterative thinning algorithm is applied by deleting border points satisfying conditions of topology preservation. On the generated skeleton, the skeleton-marching algorithm is performed to identify the basic entities and construct the skeletal graph. Basic skeletal entities are vertex, edge and loop.

For the definition of feature vectors the following shape descriptors are extracted from the voxel model: moments, geometry parameters such as volume and surface area, voxelization parameters such as voxel size, and graph parameters such as number of loops edges and nodes. For the similarity measurements the Euclidean distance of the feature vectors, as well as the distance between skeletal graphs are calculated. For the graph matching a decision-tree based algorithm developed by Messmer et al. [28] is applied.

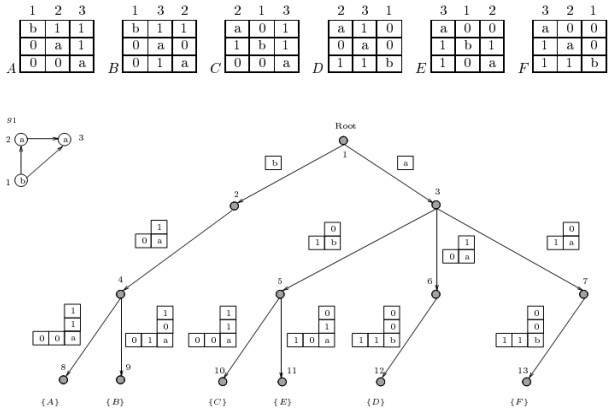


Figure 7. Decision tree for the related (above) adjacency matrix [28]

In this algorithm all graphs in a database are indexed in the form of a decision tree using the different permutations of the adjacency matrix as described in Figure 7. The space requirements are exponential, but the search requirement is sub-polynomial in the number of query graph nodes.

E. Reeb Graphs

Hilaga et al. [29] developed Multiresolutional Reeb Graphs (MRGs) for similarity measurements between 3D models. MRG describes the skeletal and topological structure of a 3D model. A reeb graph is generated using a continuous scalar function on the 3D model. In this method, geodetic distance is used because of the translation and rotation invariance of this function. Because the reeb graph might contain many nodes, a MRG is constructed as a row of reeb graphs at several levels of detail. 3D object is divided into a number of ranges using values of the scalar function. A graph node represents a connected component in a particular range, and graph edge represent connected components of the adjacent ranges that contact each other.

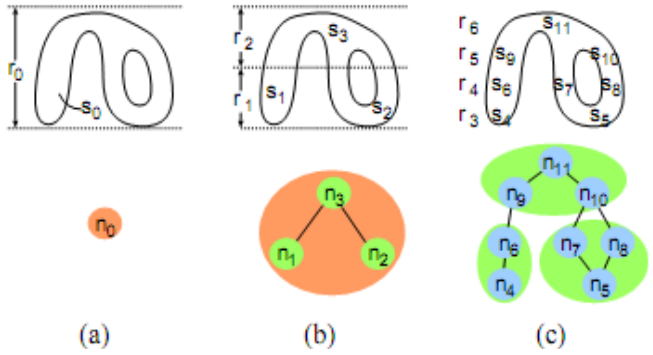


Figure 8. Multiresolutional Reeb graph [29]

The construction starts with the generation of a reeb graph having the finest resolution. Construction of the MRGs with coarser resolution is followed by merging adjacent ranges and unifying connected nodes form this ranges into one. For every node attributes are calculated and used to estimate similarity between nodes in the similarity measurements process. The calculation of node attributes is based on shape features such as area of triangles, area of whole object and certain values of continuous function. The similarity measurement is performed using a coarse-to-fine strategy, while preserving the consistency of the graph structure.

F. Evaluation of graph-based methods

The advantage of all graph-based methods is description of the topology of 3D models which is an important shape feature. In addition, representation of 3D models as topological graphs which facilitate the abstraction of these models at different levels of detail and description of local geometry at each node. [1] Other advantage of the graph-based methods is possibility of partial matching between 3D models. (Except for MRG method)

MSGs are efficient despite the large and complex graphs in the database. This method is considered insufficient for fine discrimination between 3D models, referring to the disability of capturing all properties of the adjacency matrix by the eigenvalues. [17][20]

Although skeletal graphs with using simplification of 3D, are stable to small changes in shape, but the simplification causes a loss of information affecting the discrimination power of the method. [27]

The advantage of MRG refers to the fact that geodesic distance as a continuous function is invariant against rotation and translation. Exponential computational complexity of this method has been avoided by applying coarse-to-fine strategy in the retrieval process. [29][30]

V. CONCLUSION AND FUTURE WORK

In this paper, shaped-based similarity and design retrieval methods have been discussed. Each of the discussed methods has advantages as well as disadvantages. Which method should be used in a particular application depends on the desired discrimination power of descriptor or the required efficiency of the similarity search. If only general classification of objects in a database is needed then harmonics-based or invariant-based is good choice. If partial matching between objects should be possible, then one of the graph-based methods is good choice. As a conclusion, the combination of different methods may help to achieve high discrimination power as well as efficient similarity search.

The next step will be review and discussion of the rest of the methods and classes, which decompose a shape into a signature, i.e., Statistics/probability-based methods, 3D Object Recognition-based methods, Feature Recognition (FR)-based methods, and Group Technology (GT)-based methods.

REFERENCES

[1] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani, “Shape-based searching for product lifecycle applications”, Computer-Aided Design, vol. 37, issue 13, pp. 1435-1446, November 2005

[2] G. Cybenko, A. Bhasin, and K. Cohen, “Pattern recognition of 3D CAD objects”, Smart Engineering System Design, vol. 1, pp. 1-13, 1997

[3] M. Kazhdan and T. Funkhouser, “Harmonic 3D Shape Matching”, SIGGRAPH Sketches and Applications, pp. 191-191, July 2002

[4] M. Elad, A. Tal, and S. Ar, “Content based retrieval of VRML objects -an iterative and interactive approach” Eurographics multimedia workshop; pp. 97-108, 2001

[5] M. Elad, A. Tal, and S. Ar, “Directed search in a 3D objects database using SVM”, Technical Report, HPL-2000-20(R.1), HP Laboratories Israel, 2000

[6] H. Rea, J.R. Corney, D. Clark, J. Pritchard, M. Breaks, and R. MacLeod, “Part sourcing in a global market”, Concurrent Engineering: Research and Applications, 10 (4), pp. 325-334. ISSN 1063-293X, 2002

[7] J.R. Corney, H. Rea, D. Clark, J. Pritchard, M. Breaks, and R. MacLeod, “Coarse filters for shape matching”, IEEE Computer Graphics and Applications, vol. 22, issue 3, pp. 65-74, 2002

 ISBN: 978-1-61208-162-5

[8] C. Zhang, and T. Chen, "Efficient feature extraction for 2D/3D objects in mesh representation" Proc. ICIP2001, vol.3, pp. 935-938, doi: 10.1109/ICIP.2001.958278

[9] C. Zhang, and T. Chen, "Active Learning for Information Retrieval: Using 3D Models As An Example", Technical Report AMP 01-04, pp. 01-04, Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, 2001

[10] D. Vranic, D. Saupe, and J. Richter, "Tools for 3D-object retrieval: Karhunen-Loeve transform and spherical harmonics", Proc. IEEE MMSP workshop, 2001, pp. 293-198, doi: 10.1109/MMSP.2001.962749

[11] D. Saupe and D. V. Vranic, "3D Model Retrieval With Spherical Harmonics and Moments", Proc. DAGM 2001 (editors B. Radig and S. Florczyk), Munich, Germany, pp. 392-397, September 2001

[12] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors" Proc. ACM/eurographics symposium on geometry processing, pp. 167-175, 2003

[13] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Helderman, and D. Dobkin, "A Search Engine for 3D Models", ACM Transaction on Graphics, vol. 22, pp. 83-105, 2003

[14] D. Vranic, "An Improvement of rotation invariant 3D shape descriptor based on function on concentric spheres", IEEE International Conference on Image Processing Barcelona, vol. 3, pp. 757-760, 2003

[15] D. Vranic, "3D model retrieval", Ph. D. Dissertation, 2004, University of Leipzig, Department of Computer Science

[16] P. Papadakis, I. Pratikakis, S. Perantonis, and T. Theoharis, "Efficient 3D shape matching using a concrete radialized spherical projection representation", Pattern Recognition, vol. 40, issue 9, pp. 2437-2452, 2007

[17] D. McWherter, M. Peabody, W. C. Regli, and A. Shokoufandeh, "An approach to indexing databases of graphs", Technical Reports DUMCS-01-01, Philadelphia, PA: Department of Mathematical and Computer Science, Drexel University, Juni 2001

[18] D. McWherter, M. Peabody, W. C. Regli, and A. Shokoufandeh, "Transformation invariant shape similarity comparison of solid models", ASME Design Engineering Technical Conference, American Society of Mechanical Engineers, ASME Press. DETC 2001/DFM-21191, 2001

[19] D. McWherter, M. Peabody, W. C. Regli, and A. Shokoufandeh, "Solid Model databases: techniques and empirical results". Journal of Computing and Information Science in Engineering, Vol. 1, No. 4, pp. 300-310, 2001, doi:10.1115/1.1430233

[20] M. Peabody and W. C. Regli, "Clustering techniques for databases of CAD models". Technical Report DU-MCD-01-0. Philadelphia, PA: Department of Mathematical and Computer Science, Drexel University, 2001

[21] M. El-Mehalawi and R. Miller, "A database system of mechanical components based on geometric and topological similarity. Part I: Representation", Journal of Computer-Aided Design, vol. 35 (1), pp. 83-94, 2003

[22] M. El-Mehalawi and R. Miller, "A database system of mechanical components based on geometric and topological similarity. Part II: indexing, retrieval, matching and similarity assessment", Journal of Computer-Aided Design, vol. 35, issue1, pp. 95-105, 2003

[23] H. Sundar, D. Silver, N. Gagvani, and S. Dickson, "Skeleton based shape matching and retrieval", Proceedings of Shape Modeling and Application, Korea , pp. 130 – 139, 2003

[24] N. Gagvani and D. Silver, "Parameter Controlled Volume Thinning", Graphical Models and Image Procesing, vol. 61, issue3, pp.149-164, May 1999

[25] N. Iyer, Y. Kalyanaraman, K. Lou, S. Jayanti, and K. Ramani, "A reconfigurable 3D engineering shape search system Part I: shape representation", Proc. ASME DECT 03 Computers and Information in Engineering (CIE) Conference, pp. 89-98, Chicago, IL, 2003

[26] K. Lou, N. Iyer, Y. Kalyanaraman, K. Ramani, and S. Prabhakar, "A reconfiguable 3D engineering shape search system. Part II: database indexing, retrieval and clustering", Proc. ASME DECT 03 Computers and Information in Engineering (CIE) Conference, pp. 169-178, Chicago, IL, 2003

[27] N. Iyer, K. Lou, S. Jayanti, Y. Kalyanaraman, and K. Ramani, "A Multi-Scale Hierarchical 3d Shape Representation for Similar Shape Retrieval", TMCE 2004, Lausanne, Switzerland, pp. 1117-1118, 2004

[28] B. Messmer and H. Bunke, "Subgraph isomorphism in polynomial time", Technical Report TR-IAM-95-003, 1995

[29] M. Hilaga, Y. Shinagawa, T. Kohmura, and T.L. Kunii, "Topology matching for fully automatic similarity estimation of 3D shapes", SIGGGRAPH, ACM Press, pp. 203-212, 2001

[30] D. Bespalov, W.C. Regli, and A. Shokoufandeh, "Reeb graph-based shape retrieval for CAD", Proceeding of the ASME DECT 03 Computers end Information in Engineering (CIE) Conference, pp. 229-238, Chicago, IL, 2003

 ISBN: 978-1-61208-162-5

Mining Information Retrieval Results

Significant IR parameters

Jonathan Compaoré
Adji Maïram Gueye
Institut National des Sciences Appliquées
Université de Toulouse
Toulouse, France
{jcompaor/amgueye}@etud.insa-toulouse.fr

Sébastien Déjean
Institut de Mathématique de Toulouse
Université de Toulouse
Toulouse, France
sebastien.dejean@math.univ-toulouse.fr

Josiane Mothe
Joelson Randriamparany
Institut de Recherche en Informatique de Toulouse
Université de Toulouse
Toulouse, France
{mothe/randriamparany}@irit.fr

***Abstract*—This paper presents the results of mining a large set of information retrieval results. The objective of this study is to determine which parameters significantly affect search engine performance. We focus on the main features of information retrieval: indexing parameters and search models. Statistical analysis identifies the retrieval model as the most important parameter to be tuned to improve the performance of an information retrieval system. We also show that the significant parameters depend on the difficulty of the topic.**

Keywords-information retrieval; data mining; parameter analysis; performance prediction.

I. Introduction

Information retrieval systems are generally evaluated considering their effectiveness. While evaluated in international campaigns such as TREC [1] and CLEF [2], etc., consider the Cranfield evaluation model [3]. This model consists of a collection of documents on which a query is evaluated, a set of queries and the list of relevant documents for each of these queries. Evaluation frameworks also imply performance measurements that will be helpful to compare systems.

Evaluation campaigns have contributed a lot to the Information Retrieval (IR) field [4][5]. They clearly contribute to the definition of new models and new processes such as blind relevance feedback for example [6]. From the published performance results it is possible to know which systems perform best on the set of topics suggested by the evaluation program. In addition, from the system description published onin the associated papers, it is generally possible to know in detail the type of systems used and even the various parameters used in a specific experiment.

On the other hand, what the results hide is the individual contribution of a given component. Indeed from these experiments, it is not possible to know what the impact of the indexing is, nor a given parameter. The impact of parameters in models is evaluated when defining the model. For example, Zhai and Lafferty [7] present the precision when various parameters of smoothing methods for language modeling vary. In this paper our goal is different, we aim at discovering if some parameters are significantly correlated with some performance measures. To discover such information, we use a platform that implements various indexing schemes and search models and that allow one to modify parameter values. Then we process the same topics using these various system configurations and evaluate the results in terms of effectiveness. We then analyze the resulting data with the aim of finding associations between system parameters and performance measures.

The rest the paper is organized as follows. In Section 2, we present related works. In Section 3, we describe the platform we use and the way we obtain the data to be analyzed. In Section 4, we present the preliminary results we obtained using one performance measure MAP (Mean Average Precision). Section 5 presents conclusions and outlines directions for future work.

II. Related works

An IR process is composed of various components. A first component is indexing which is done off line, prior to any querying. Indexing aims at deciding which terms will be used to represent document content and to match the document and the query. The indexing unit, stop-word list, stemming algorithm, and term weighting function are among the parameters of indexing. When queried, the IR system has then to match the query with the document. This is done through a similarity function or a ranking function generally based on content similarity (term comparison). Systems vary according to the model used: vector space model [8], Probabilistic model [9][10], LSI [11], language modeling [12]; each model has parameters that can be modified.

Generally speaking, related work analyzes one component or a few components of the retrieval process in terms of its influence on the effectiveness, considering for example MAP.

Kompaoré *et al.* [13], for example, focus on the indexing part. They analyze three types of indexing units applied on French test collections: lemmas, truncated terms and single words. Considering the 284 used in the CLEF French track,

 ISBN: 978-1-61208-162-5

they show that the best results were obtained while considering lemmas. Lifchitz *et al.* [14] analyze the effect of parameters on the LSA model, a model which is based on singular decomposition of the term/document matrix resulting from indexing. They show that normalizations of documents and term frequencies have a negative effect on the results. They also conclude that the optimal truncation (number of dimensions) of the semantic space and the stop word list play a major role.

The "reliable information access" workshop [15][16] focused on variability and analyzed both system and topic variability factors on TREC collections when query expansion is used. Seven systems were used, all using blind relevance feedback. System variability was studied through the different systems by tuning different system parameters and query variability was studied using different query reformulation strategies (different numbers of added terms and documents). Several classes of topic failure were drawn manually, but no indications were given on how to automatically assign a topic to a category.

Bank *et al.* [17] reports various data analysis methods and how they can be used to analyze TREC data. They consider analysis of variance, cluster analyzes, rank correlations, beadplots, multidimensional scaling, and item response analysis. When considering analysis of variance, they consider two effects only: topic and system and one performance measure (average precision). They also used cluster analysis to cluster systems according to MAP they obtained. Even if this preliminary study concluded that none of these methods has yielded any substantial new insights; more recent work [18] has shown that clustering can be used in the case of repeated queries.

III. Generating data

A. TREC data

Since we want to evaluate individual parameters of the search process, it is compulsory to consider the same collection (information needs, documents on which the search is carried out and relevance judgments). International experimental environments, such as TREC, provide such a framework.

The *ad hoc* task was introduced in the earlier years of TREC in 1991. It simulates a traditional IR task for which a user queries the system. The system retrieves a ranked list of documents that answer this query from a static set of documents.

We work in this paper with data used two consecutive years in TREC-7 and TREC-8 collections. The document collection consists of disks 4 and 5 which corresponds to 528 155 documents. A total of 100 topics are used here, topics 351-400 used in TREC-7 and topics 401-450 used in TREC-8. Details on the set of documents and topics can be found in [1].

An example of a TREC topic is presented in Figure 1. In addition to its identifier, a topic is composed of a title, a descriptive and a narrative part. Title only can be used to build a query to be submitted to a system. It is also possible to build the query from other topic part combinations.

<num> 396
<title> sick building syndrome
<desc> Identify documents that discuss sick building syndrome or building-related illnesses
<narr> A relevant document would contain any data that refers to the sick building or building-related illnesses, including illnesses caused by asbestos, air conditioning, pollution controls. Work-related illnesses not caused by the building, such as carpal tunnel syndrome, are not relevant

Figure 1. Example of TREC topic (#396).

B. Terrier platform

Terrier is an information retrieval platform that implements state-of-the-art indexing and retrieval functionalities.

1) Topics

One of the parameters of a search is the topic used. In our experiments, we consider 100 topics, numbered 351 to 450.

From these topics, queries are built using an indexing process (see below). From a topic, three types of queries are built: using title (`T`) only, title and descriptive (`TD`), title, descriptive and narrative (`TDN`), referred as variable *Field*.

2) Indexing

Document indexing is used to extract indexing terms from document contents. Usually, stop-words are removed and remaining terms are stemmed in order to conflate the variant of words into a single form. Indexing terms are weighted in order to reflect their descriptive power.

While indexing documents, they can be cut into several chunks of text. This process has an impact on the term weights. The number of blocs is a parameter (variable *Bloc*). The value is `1` when any document is considered as a unit; the two other values we used are `5` and `10`.

Regarding the weighting schema, it is possible to consider the inverse document frequency. For these reason the parameter *Idf* is set either to 0 (`FALSE`) or to 1 (`TRUE`).

3) Retrieval models

Nine models are implemented in Terrier. We use each of them (variable *Model*): `BB2c1`, `BM25b0.5`, `DFRBM25c1.0`, `IFB2c1.0`, `InexpB2c1.0`, `InexpC2c1.0`, `InL2c1.0`, `PL2c1.0`, `TFIDF`. Details on these models can be obtained at Terrier web site [19].

4) Query expansion

Query reformulation is used in order to improve the initial query so that it can retrieve more relevant documents. Blind relevant feedback is used [6].

We used the three models implemented at the Terrier: `Bo1bfree`, `Bo2bfree`, `KLbfree` (parameter *Ref*).

The number of documents used during query reformulation is a parameter (*DocNb*). It varies in {`0`, `3`, `10`, `50`, `100`, `200`}. The number of documents in which the

terms must occur to be considered as relevant to be used in the expended query is a parameter (*qe_md*). Its value is either `0` or `2`. The number of terms added to the initial query is the last parameter of query expansion; its value is either `0` or `10` (*qe_t*).

Some variables appear to be redundant. For example, *qe_md* and *qe_t* are redundant since as soon as there is an expansion, there will be 10 terms added and they should occurs at least in 2 documents. The variables are presented in Table 1.

TABLE 1. PARAMETERS AND VALUES USED FOR A SEARCH.

Parameters	Meaning	Values
Top	Topic number	351, …, 450
Field	Topic field	`T`; `T+D`; `T+D+N`
Bloc	Size of the indexing bloc	`1`, `5`, `10`
Idf	Inverse document frequency	`FALSE`, `TRUE`
Ref	Query reformulation	`None`, `Bo1bfree`, `Bo2bfree`, `KLbfree`
Model	Retrieval model	`BB2c1`, `BM25b0.5`, `DFRBM25c1.0`, `IFB2c1.0`, `InexpB2c1.0`, `InexpC2c1.0`, `InL2c1.0`, `PL2c1.0`, `TFIDF`
DocNb	Number of documents (reformulation)	`0`, `3`, `10`, `50`, `100`, `200`
qe_md	Minimum number of documents in which the term should appear to be used in the query epension	`0`, `2`
qe_t	Number of terms used in the query epension	`0`, `1`

A combination of these parameters leads to a run that can be evaluated using performance measures.

C. Performance measures

We use the TREC software trec_eval to evaluate each individual run. Measures are computed for each topic. The version 8.1. of trec_eval [20] that we used computes 135 measures. Baccini *et al.* [21] have shown that many performance measures are redundant and that it is possible to keep 6 representative measures that will cover the various aspects of IR evaluation. The remainder of the analysis focus on only one performance measure (MAP) for illustration purpose.

D. Data to analyze

We generate a matrix that is composed of:

- 98650 objects (lines of the matrix). Each line corresponds to one topic processed by a chain of modules (indexing, search) and evaluated according to various performance measures;
- 8 variables (columns of the matrix) that consist in 7 non redundant module parameters (see Table 1 minus *qe_md* and *qe_t*) and 1 performance measure (MAP).

The value in a cell corresponds to the characteristic of the object for the corresponding variable. When it is a system parameter the cell contains the value of the parameter; when it is a performance measure, the cell contains the result of the evaluation.

IV. STATISTICAL ANALYSIS

We aim at identifying which parameters have a significant influence on the performance of the system. To address this question, we first performed an Analysis of Variance (ANOVA) to explain MAP according to each parameter separately. Then using Classification and Regression Trees (CART, [22]) every parameter is jointly analyzed. These methods were applied in three frameworks:

- One global analysis: every topics were considered;
- Two restricted analysis: considering only the easiest (resp. hardest) topics. Easiest (resp. hardest) topics are the ones for which the average AP over the systems is the highest (resp. lowest).

A. Analysis of Variance (ANOVA)

ANOVA tests whether or not the mean of several groups are all equal. In our context, this will result in testing whether the MAP is significantly different when considering various configurations of one parameter.

1) Global analysis

The results of the global analysis are summarized in Figure 2. Parallel boxplots corresponding to the various categories are displayed for each parameter. First, it appears that *Bloc* and *Idf* (white boxplots indicates non significant effect) have no influence on the performance. Considering Field, MAP is lower when using only the title (`T`) of the topic; results are nearly equivalent for `TD` and `TDN`. For *Ref*, the absence of reformulation (`NONE`) seems to be detrimental in relation to one system for query reformulation whatever it is (`Bo1bfree`, `Bo2bfree` or `Klbfree`).

The main comment regarding the retrieval model (variable *Model*) is the bad behavior of `BM25b0.5`, and to a lesser extent of `PL2c1.0`. Then, *DocNb* (the number of documents used in query reformulation) provides the best results with `3` or `10` documents. Then, the MAP decreases as the number of documents used for the reformulation increases.

These results outline phenomenon visible at a global scale considering all the topics. We wanted to see if these results hold for two particular sub-sets of topics: the easiest and the hardest.

2) Analysis of easy and hard topics

Most of the work in IR considers the result globally, averaging the results over a set of queries. On the other hand, some works have shown that system results (e.g. AP) is query dependent [18]. Finally, some studies have focused on hard topics, trying to find ways to handle them better [16]. Finally, Bigot *et al.* [18] show that choosing the best system for individual queries improves results differently according to query difficulty. For that reasons, we decided to consider two types of topics: the hard and easy topics and to analyze the behavior of the parameters according to these

 ISBN: 978-1-61208-162-5

two topic sets.

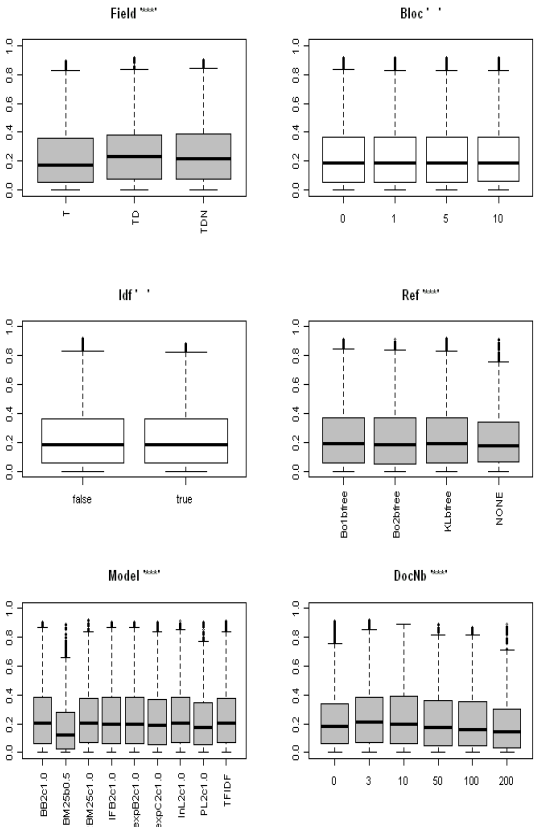


Figure 2. Boxplots representing MAP according to the different levels of each parameter (Field – 3 levels, Bloc – 4 levels, Idf – 2 levels, Ref – 4 levels, Model – 9 levels, DocNb – 6 levels). The symbol near the title indicates the p-value of the test according to the code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Grey boxplots highlights cases where the parameter is highly significant in the ANOVA model (p-value < 0.001).

The "easy" topics are the ones for which the average AP over systems is the highest. Considering the analyzed data, the corresponding AP is higher than 0.45 (considering this value, there is a gap between topics). There are 13 topics in the easy topic set. In the same way, hard topics correspond to topics for which the mean of AP over systems is the lowest. AP is lower than 0.045. There are 19 topics in the hard topic set.

The results of the restricted analysis highlights more heterogeneity in the behavior of MAP according to parameters. For the easy topics, 4 of the 6 parameters have a statistically significant influence on the MAP.

The most visible phenomenon is the very particular profile of the *Model* parameter: using `BM25b0.5` significantly deteriorates MAP. This is also the case for the hard topics but it also appears that some retrieval models, such as `ExpC2c1.0` or `BB2c1.0` are more appropriate to improve MAP.

The influence of the field used when indexing topics (`T`, `TD` or `TDN`) seems to have more impact for hard topics. The dispersion of the MAP values is also higher when considering hard topics. Indeed generally speaking, when considering hard topics there is a larger range of values than when considering easy topics. That is to say, there is no failure when considering easy topics whereas there are some good results for hard topics.

Interesting enough, the impact of the number of documents when using query reformulation is not the same over the two restricted analysis: while `3` or `10` documents seems to be the best choice for easy topics, `0` to `3` is more appropriate for hard topics.

Hard topics also exhibit two other configurations to use to improve MAP: the query reformulation proposed by `Bo1bfree` and the *Idf* set to `FALSE`. However, let's note that even if the MAP is increased it remains relatively low.

B. Classification and Regression Trees

In the previous section, we analyzed the influence of each parameter of the IR process on the results. In this section, we try to sketch a strategy on parameter tuning during the IR process in order to maximize the performances (according to MAP) of a search. To address this question, we also want to deal with the parameters simultaneously and to consider potential interactions between them. This purpose can be achieved by using CART [22]. The main idea in CART lies in the construction of a decision tree by splitting successively the observations depending on the values of the dependent variables which is numeric for regression and categorical for classification. Details can be found in the original article by Breiman *et al.* [22] or in a more recent review [23]. We opted for the classification version of CART in order to identify the most important parameters to be tuned to obtain better results without quantifying the resulting value. Implementation was performed using the *rpart* package [24] of the R software [25].

1) Data

Considering CART in the classification framework required MAP values to be converted into qualitative information. We used the quartiles to divide the range of the MAP into 3 classes. Then, we code the values according to which interval they fall in. We use three tags: "*Bad*" (MAP lower than the first quartile), "*Average*" (MAP between the first and the third quartile) and "*Good*" (MAP greater than the third quartile). Table 2 reports the values of MAP corresponding to these tags according to the type of topics used. Indeed, we considered three sets of queries, like in subsection IV A. Global data consists of the all set of topics, easiest topics and hardest topics corresponding to the sets defined in Section 4.

 ISBN: 978-1-61208-162-5

TABLE 2. MAP VALUES DEPENDING ON THE TOPIC TYPES AND TAGS.

Tag	Global	Easiest	Hardest
Bad	<0.057	<0.48	<0.005
Average	0.057 ≤MAP≤0.37	0.48 ≤MAP≤0.69	0.005 ≤MAP≤0.035
Good	>0.37 (max=0.92)	>0.69 (max=0.92)	>0.035 (max=0.21)

2) Results

Figure 3 presents the results when considering the global set of topics. The categories of each qualitative variables are coded by letters. For instance, the 9 categories for the variable *Model* are coded from *"a"* to "*i*". The tree indicates that the variable *Model* is the most important (first node on top of the tree) to classify the runs. When *Model* is not "b" (BM25b0.5), most of runs get "*Average"* MAP (right child node). When *Model* is "b" (`BM25b0.5`), the next most important variable is *Field*. If *Field* is `TDN` 'label "c"), results can be "*Average*" else *Ref* becomes important to go on the classification and so on...

The results obtained in Figure 3 are consistent with the information provided by the boxplots (Figure 2). The worst results are obtained with `BM25b0.5` as the second boxplot for *Model* is clearly moved downward. The fact that the class "*Good"* does not appear in the tree is also consistent with Figure 2 as none boxplot appears above the others.

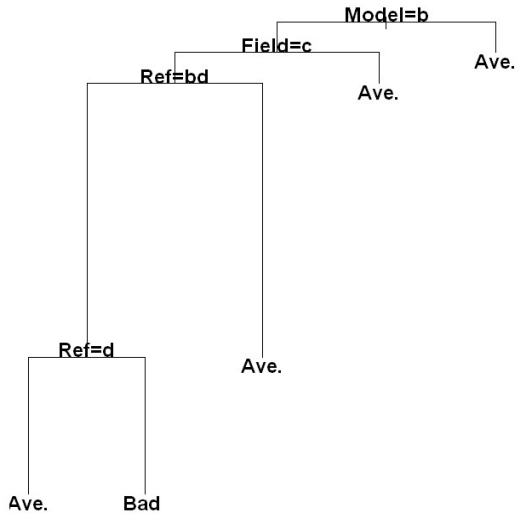


Figure 3. CART obtained with the global set of topics. The categories of each parameter are coded with letters according to the order given in Table 1.

Regarding the easiest topics (Figure 4), the parameter *Model* is still the most important in classifying the 3 categories of MAP. In addition to `BM25b0.5`, `PL2c1.0` (coded as "h") is also considered as a bad configuration. The next most important parameter is *DocNb* which provides lower values of MAP ("*Bad*" label) with `100` ("e") and `200` ("f") documents used.

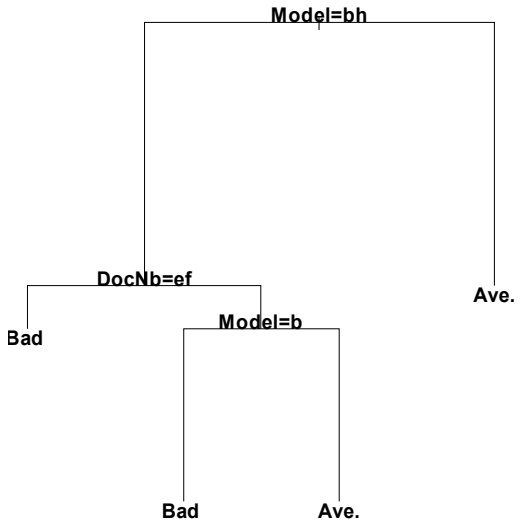


Figure 4. CART obtained with the easiest topics.

For the hardest topics (Figure 5), the structure of the tree appears more complicated although the same pruning parameters was used. Four parameters are involved: *DocNb*, *Ref*, *Model* and *Idf*. Surprisingly, *Field* does not appear in the tree. This is certainly due to potential interactions between parameters that are not taken into account with univariate ANOVAs.

VI. CONCLUSION AND FUTURE WORKS

The analysis we performed clearly confirms that some parameters produce significant changes in the performance of information retrieval systems. It also indicates that these changes are different when considering various topics characterized by unequal difficulty and provides clues to tune the parameters in order to improve the performance.

The analysis we conduced does not permit to predict performance measures but indicates the parameters that have the higher influence on the results. One important result of this study is that parameters depend on the difficulty of the topic.

This work has to be validated considering other performance measures and a more systematic procedure to characterize groups of topics.

VI. ACKNOWLEGMENTS

This work was supported in part by the ANR Agence Nationale de la Recherche (CAAS project), by the Région Midi-Pyrénées (project #10008510) and by the federative research structure FREMIT.

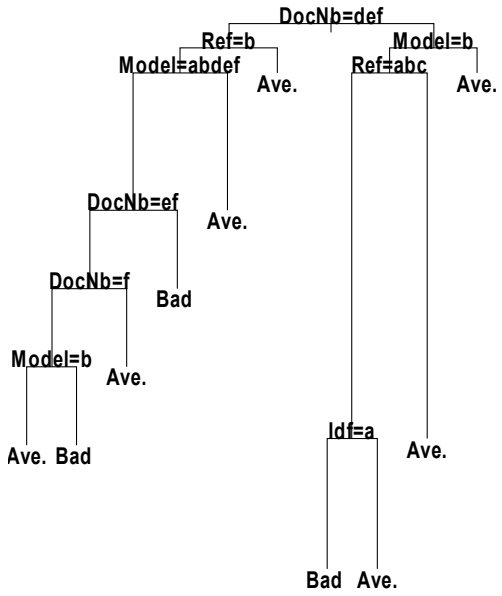


Figure 5. CART obtained with the hardest topics.

References

[1] TREC Text REtrieval Conference at http://trec.nist.gov <retrieved: 08,2011>

[2] CLEF Cross-Language Evaluation Forum at http://clef-campaign.org <retrieved: 08,2011>

[3] C. W. Cleverdon, J. Mills, and E. M. Keen, "Factors determining the performance of indexing systems. Cranfield", UK: Aslib Cranfield Research Project, College of Aeronautics, 1966, Volume 1:Design; Volume 2: Results.

[4] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. Voorhees, "Report on the SIGIR 2009 Workshop on the future of IR evaluation", ACM SIGIR Forum, Vol. 43 Issue 2, 2009, pp. 13-23.

[5] S. Robertson, "Richer Theories, Richer Experiments, The Future of IR Evaluation Workshop", Inter. ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, 4.

[6] D. Evans and R. Lefferts, "Design and evaluation of the clarit-trec-2 system", 2nd Text Retrieval Conference, NIST Special Publication 500-215, 1994, pp. 137-150.

[7] C. Zhai, J. Lafferty, "A study of smoothing methods for language models applied to information retrieval", ACM Transactions on Information Systems, Vol. 22 Issue 2, 2004, pp. 179 – 214.

[8] G. Salton. "The Smart Retrieval System", Prentice Hall, Englewood Cliffs, NJ, 1971.

[9] S. E. Robertson and K.S. Jones, "Relevance weighting of search terms". Journal of the American Society for Information Sciences, Vol. 27, Issue 3, 1976, pp. 129-146.

[10] S. E. Robertson and S. Walker, "Some simple approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval", Proc. ACM SIGIR conference on Research and development in information retrieval, 1994, pp. 232-241.

[11] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis". JASIST. Vol. 41, Issue 6, 1990, pp. 391-407.

[12] J. M. Ponté and B. Croft, "A langage modeling approach to information retrieval", Proc. ACM SIGIR, Conference and Research and Development in Information Retrieval, 1998, pp. 275-281.

[13] N.D. Kompaoré, J. Mothe, and L. Tanguy, "Combining indexing methods and query sizes in information retrieval in French", Proc. International Conference on Enterprise Information Systems (ICEIS), 2008, pp. 149-154.

[14] A. Lifchitz, S. Jhean-Larose, and G. Denhière, "Effect of tuned parameters on an LSA multiple choice questions answering model", Behavior Research Methods, Vol. 41 Issue 4, 2008, pp. 1201-1209.

[15] D. Harman and C. Buckley, "The NRRC reliable information access (RIA) workshop", Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 528-529.

[16] D. Harman and C. Buckley, "Overview of the Reliable Information Access Workshop", Information Retrieval, Vol. 12, Issue 6, 2009, pp. 615-641.

[17] D. Banks, P. Over, and N.F. Zhang. "Blind Men and Elephants: Six Approaches to TREC data", Information Retrieval, Vol. 1, Issue 1-2, 1999, pp. 7-34.

[18] A. Bigot, C. Chrisment, T. Dkaki, G. Hubert, and J. Mothe, "Fusing Different Information Retrieval Systems According to Query Topics - A Study Based on Correlation in Information Retrieval Systems and Query Topics ", DOI: 10.1007/s10791-011-9169-5 (to appear).

[19] Terrier web site, http://terrier.org/, <retrieved: 08,2011>.

[20] Trec_eval, http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval_video/README version from the 24th july 2006, <retrieved: 08,2011>.

[21] A. Baccini, S. Déjean, L. Lafage, and J. Mothe. "How many performance measures to evaluate Information Retrieval Systems?" Knowledge and Information Systems, DOI 10.1007/s10115-011-0391-7, 2011 .

[22] L. Breiman, J.H. Friedman, R. Olshen, and C.J. Stone. Classification and Regression Trees. 1984, Chapman & Hall/CRC.

[23] W.-Y. Loh, "Classification and regression tree methods". In Ruggeri, F., Kenett, R., and Faltin, F. W., editors, Encyclopedia of Statistics in Quality and Reliability, 2008, pp. 315–323. Wiley, Chichester, UK.

[24] T.M. Therneau and B. Atkinson. R port by Brian Ripley. (2010). rpart: Recursive Partitioning. R package version 3.1-48. http://CRAN.R-project.org/package=rpart <retrieved: 08,2011>

[25] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/<retrieved: 08,2011>.

 ISBN: 978-1-61208-162-5

Mining Literal Correlation Rules from Itemsets

Alain Casali
Laboratoire d'Informatique Fondamentale de Marseille (LIF),
CNRS UMR 6166, Aix Marseille Universités, IUT d'Aix en Provence
Aix en Provence, France
alain.casali@lif.univ-mrs.fr

Christian Ernst
Ecole des Mines de St Etienne
CMP - Site Georges Charpak
Gardanne, France
ernst@emse.fr

***Abstract*—Nowadays, data mining tools are becoming more and more popular to extract knowledge from a huge volume of data. In this paper, our aim is to extract Literal Correlation Rules: Correlation Rules admitting literal patterns given a set of items and a binary relation. If a pattern represents a valid Correlation Rule, then any literal belonging to its Canonical Base represents a valid Literal Correlation Rule. Moreover, in order to highlight only relevant Literal Correlation Rules, we add a pruning step based on a support threshold. To extract such rules, we modify the LHS-CHI2 Algorithm and perform some experiments.**

***Keywords*-Data Mining; Chi^2 Correlation Statistic; Literal Pattern.**

I. INTRODUCTION AND MOTIVATION

An important field in data mining is the discovery of links between values (items) in a binary relation in reasonable response times. Agrawal *et al.* [1] introduce levelwise algorithms in order to compute association rules. Those latter express directional links ($X \rightarrow Y$ for example), based on the support/confidence platform. From this problem, three sub-problems are particularly interesting.

The first one is an adaptation of the supervised classification [2], [3]. Instead of making unsupervised classification, the authors consider the presence of several target attributes. They only consider associations in which the right hand side of the rule contains at least a value of a target attribute. Moreover, they apply rules (specific to the method) allowing to predict to which class belongs an unknown pattern.

The second one is the introduction of literal patterns by Wu *et al.* [4]. The authors compute positive and/or negative association rules, such as $\neg X \rightarrow Y$. To generate the rules, they still use the support/confidence platform by redefining the support of a literal.

In the third one, Brin *et al.* [5] propose the extraction of Correlation Rules, where the platform is no longer based on the support nor the confidence of the rules, but on the Chi-Squared statistical measure, written χ^2. The use of χ^2 is well-suited for several reasons: (i) It is a more significant measure in a statistical way than an association rule; (ii) The measure takes into account not only the presence but also the absence of the items; (iii) The measure is non-directional, and can thus highlight more complex existing links than a "*simple*" implication.

Unlike Association Rules, a Correlation Rule is not represented by an implication but by a set of items for which the value of the χ^2 function is larger or equal than a given threshold, noted $MinCor$.

Since the crucial problem, when computing correlation rules, is the memory usage required by levelwise algorithms, [5] compute only correlations between two values of a binary relation. In [6], we introduce the LHS-CHI2 algorithm. The objective is to compute Decisional Correlation Rules (Correlation Rules which contain, at least, a value of a target attribute). To achieve such an objective, we change the strategy of the browsing search space. We use a lectic order strategy [7] instead of a levelwise one. Since we could not find a function f linking the correlation rate, related to a pattern X, with any of its supersets, we introduce the concept of Contingency Vector, another representation of a Contingency Table based on partitions. Using this concept, we found a function linking the contingency vector of a pattern X with the ones of any of its supersets. Using the LHS-CHI2 algorithm, we have a gain of execution times between 30% and 70% compared to a levelwise algorithm.

Moreover, when applying Advanced Process Control approaches in semiconductor manufacturing, it is important to highlight correlations between parameters related to production, in order to rectify possible drifts of the associated processes. Within this framework, and in collaboration with STMicroelectronics and ATMEL, our previous work [6] focuses on the detection of the main parameters having an impact on the yield. We extract correlations between the values of some columns and those of a target column (a particular column of the file, the yield).

In this paper, we focus on finding out correlations with literal patterns. Such information are important either to point out fault detection, and to detect what parameters do not have an impact on the yield (while they should have). To solve this problem, we introduce Literal Correlation Rule and Literal Decision Correlation Rule concepts. The former is a Correlation Rule admitting literal patterns, and the latter is a restriction of Literal Correlation Rule containing, at least, one value of a target column. In order to compute

 ISBN: 978-1-61208-162-5

those rules:

1) We propose a new formula to compute the χ^2 over literal patterns;
2) We show that the χ^2 value for a pattern is equal to the one of "some" literals;
3) We propose a new constraint in order to highlight relevant Literal (Decision) Correlation Rules;
4) We modify the LHS-CHI2 algorithm to take into account these new constraints.

Finally, we carry out experiments on relations provided by the above mentioned manufacturers.

The paper is organized as follows: in Section II, the bases of Literal Patterns and of (Decision) Correlation Rules are recalled. Section III describes the main contribution of the paper. Experiments are detailed in Section IV. As a conclusion, we summarize our contribution and outline some research perspectives.

II. RELATED WORK

In this paper, we use the following notations: let $\mathcal{R}$ be the set of all 1-items and r a binary database relation over $\mathcal{R}$. In our context, $\mathcal{R}$ can be divided into two distinct sets, noted $\mathcal{I}$ and $\mathcal{T}$. $\mathcal{I}$ represents the values of the binary relation used for criteria analysis, and $\mathcal{T}$ is a target attribute. In this section, the concepts of Literal Patterns and Correlation Rules are first recalled.

A. Literal Patterns

Let X, Y be two subsets of $\mathcal{R}$. A literal is a pattern $X\overline{Y}$ in which X is also called the positive part and $\overline{Y}$ the negative part. Literal patterns can be used to extend the well known association rules mining problem: The goal is to obtain new semantics. In a basket market analysis context, the rule $X \rightarrow W$ symbolizes the probability to buy W if one bought X. Using literals, we can extract rules such as $X\overline{Y} \rightarrow W$. This rule materializes the probability to buy W if one bought X but no 1-item (items with cardinality 1) of Y. In [4], to compute rules with literal patterns, Wu *et al.* always use the support-confidence platform by redefining the support of a literal: The number of transactions of the binary relation including X and containing no 1-item of Y.

Example 1: The relation example r given in Table I is used to illustrate the introduced concepts. In this relation, BC and $\overline{B}C$ patterns have a support equal to 4 and 0 respectively. The association rules $B \rightarrow C$ and $\overline{B} \rightarrow C$ have a confidence equal to 1/2 and 0 respectively. This means that half of the transactions including pattern B also contains pattern C and we can not find a transaction which does not contain B and which includes C.

The Canonical Base of a pattern X groups all the possible combinations of literals $Y\overline{Z}$ such that the union between the positive and the negative parts is X, and there is no 1-item in common between those two parts. More precisely, the Canonical Base can be defined as follows:

Table I
RELATION EXAMPLE r.

Tid	$\mathcal{I}$	$\mathcal{T}$
1	BCF	G
2	BCF	G
3	DF	G
4	F	G
5	BC	H
6	BC	-
7	BD	-
8	B	-
9	BF	-
10	BF	-

Definition 1 (Canonical Base): Let $X \subseteq \mathcal{R}$ be a pattern, we denote by $\mathbb{P}(X)$ the Canonical Base associated to X. This set is defined as follows: $\mathbb{P}(X) = \{Y\overline{Z}$ such that $X = Y \cup Z$ and $Y \cap Z = \emptyset\} = \{Y\overline{Z}$ such that $Y \subseteq X$ and $Z = X \backslash Y\}$.
By extension, we can define the Canonical Base of a literal $Y\overline{Z}$ as follows: $\mathbb{P}(Y\overline{Z}) = \{Y_1\overline{Z_1}$ such that $Y_1 \subseteq YZ$ and $Z_1 = YZ \backslash Y_1\} = \mathbb{P}(YZ)$.

Example 2: The Canonical Base associated with $X = \{A, B, C\}$ contains the following elements: $\{ABC, AB\overline{C}, AC\overline{B}, BC\overline{A}, A\overline{BC}, B\overline{AC}, C\overline{AB}, \overline{ABC}\}$.

The following property expresses that, if we take two literals belonging to the same Canonical Base, then their associated Canonical Bases are the same.

Property 1: Let X be a pattern and $Y_1\overline{Z_1}, Y_2\overline{Z_2}$ two literal patterns belonging to its Canonical Base. We have: $\mathbb{P}(X) = \mathbb{P}(Y_1\overline{Z_1}) = \mathbb{P}(Y_2\overline{Z_2})$.

B. Correlation Rules and Decision Correlation Rules

In [5], Brin *et al.* propose the extraction of correlation rules. The platform is no longer based on the support nor the confidence of the rules, but on the χ^2 statistical measure. The formula to compute the χ^2 for a pattern X is:

$$\chi^2(X) = \sum_{Y\overline{Z} \in \mathbb{P}(X)} \frac{(Supp(Y\overline{Z}) - E(Y\overline{Z}))^2}{E(Y\overline{Z})} \qquad (1)$$

Such a computation requires (i) the support, and (ii) the expectation value (or average) of all literals belonging to $\mathbb{P}(X)$. The expectation value of a literal $Y\overline{Z}$ measures the theoretical frequency in case of independence of all 1-items included in $Y\overline{Z}$, see Formula (2).

$$E(Y\overline{Z}) = |r| * \prod_{y \in Y} \frac{Supp(y)}{|r|} * \prod_{z \in Z} \frac{Supp(\overline{z})}{|r|} \qquad (2)$$

Each support of each literal belonging to the Canonical Base associated to X is stored in a table called Contingency Table. Thus, for a given pattern X, its contingency table, noted $CT(X)$, contains exactly $2^{|X|}$ cells.

In our context, there is a single degree of freedom between the items. A table giving the centile values with regard to the

 ISBN: 978-1-61208-162-5

χ^2 value for X can be used in order to obtain the correlation rate for X [8].

Example 3: With the relation Example r given in Table I, Table II shows the contingency table of pattern BC.

Table II
CONTINGENCY TABLE OF PATTERN BC.

	B	$\overline{B}$	$\sum_{row}$
C	4	0	4
$\overline{C}$	4	2	6
$\sum_{column}$	8	2	10

Thus, $\chi^2(BC) \simeq 0.28$, which corresponds to a correlation rate of about 45%.

Unlike association rules, a correlation rule is not represented by an implication but by the patterns for which the value of the χ^2 function is larger than or equal to a given threshold.

Definition 2 (Correlation Rule): Let $MinCor$ be a threshold (≥ 0), and $X \subseteq \mathcal{R}$ a pattern. If the value for the χ^2 function for X is larger than or equal to $MinCor$, then this pattern represents a valid Correlation Rule.

In addition to the previous constraint, many authors have proposed some criteria to evaluate whether a Correlation Rule is semantically valid [9]:

1) As the χ^2 computation has no significance for a 1-item, we only examine patterns of cardinality larger than or equal to two;
2) Since the χ^2 function is an increasing function, we impose a maximum cardinality, noted *MaxCard*, on the number of patterns to examine;
3) The Cochran criterion: All literal patterns of a contingency table must have an expectation value different from zero and 80% of them must have a support larger than 5% of the whole population. This criterion has been generalized by Brin et al. [5] as follows: $MinPerc$ of the literal patterns of a contingency table must have a support larger than $MinSupCT$, where $MinPerc$ and $MinSupCT$ are thresholds specified by the end-user.

Example 4: Let $MinCor = 0.25$, then the correlation rule materialized by the BC pattern is valid ($\chi^2(BC) \simeq 0.28$). However, the correlation rule represented by the BH pattern is not valid ($\chi^2(BH) \simeq 0.1$).

The crucial problem, when computing correlation rules, is the memory requirement by levelwise algorithms. For a pattern X, the computation of the χ^2 function is based on a contingency table including $2^{|X|}$ cells. Thus, at level i, C_n^i candidates (where n is the number of values of r) have to be generated and stored, in the worst case scenario, as well as the associated contingency tables. With cells encoded over 2 bytes, corresponding storage space requires 2.5 GB of memory at the 3rd level, and 1.3 TB at the 4th level.

This is why we have changed the browsing search space strategy in [6]. Instead of using a lewelwise algorithm, our algorithm, called LHS-CHI2, browse the search space according to the lectic order [7]. It is based on:

1) The LS algorithm [10]. This algorithm allows the browsing of the powerset lattice using a balanced tree;
2) Contingency vectors, another representation of the contingency tables based on bit vectors;
3) A proposition which links the contingency vector of a pattern X with the ones of its immediate successors, "*i.e.*" contingency vectors of patterns $X \cup y, \forall y \in \mathcal{R} \backslash X$;
4) A pruning step based on the positive border [11], noted BD^+.

Still in order to limit the browsing search space, whatever the browsing strategy used, we only consider Correlation Rules which have a value belonging to the set $\mathcal{T}$.

Definition 3 (Decision Correlation Rule): A Decision Correlation Rule is a Correlation Rule which contains at least one value of the target attribute $\mathcal{T}$.

Using all the constraints mentioned above, it results a gain of time between 30% and 80% using our strategy than using a levewise one. The pseudo-code of the LHS-CHI2 Algorithm is given below. The pseudo-code of the procedure CREATE_CV can be found in [6]. The predicate *CtPerc* expresses the satisfiability of the Cochran criterion. The first call to LHS-CHI2 is made with $X = \mathcal{I}$ and $Y = \emptyset$.

If we want to extract all the Correlation Rules, and not only the Decision Correlation Rules, we have to prune the test "$\exists t \in \mathcal{T} : t \in X$" from line 1 of the LHS-CHI2 Algorithm.

Algorithm 1: LHS-CHI2 Algorithm.

input : X and Y two patterns
output: $\{Z \subseteq X$ such that $\chi^2(Z) \geq MinCor\}$

1 **if** $Y = \emptyset$ ***and*** $\exists t \in \mathcal{T} : t \in X$ ***and*** $|X| \geq 2$ ***and*** $\chi^2(X) \geq MinCor$ **then**
2 | **Output** X, $\chi^2(X)$
3 **end**
4 $A := max(Y)$;
5 $Y := Y \backslash \{A\}$;
6 LHS-CHI2(X,Y) ;
7 $Z := X \cup \{A\}$;
8 **if** $\forall z \in Z, \exists W \in BD^+ : \{Z \backslash z\} \subseteq W$ **then**
9 | CV(Z) := CREATE_CV(VC(X),Tid(A)) ;
10 | **if** $|Z| \leq MaxCard$ ***and*** $CtPerc(CV(Z), MinPerc, MinSupCT)$ **then**
11 | | $BD^+ := max_{\subseteq}(BD^+ \cup Z)$;
12 | | LHS-CHI2(Z,Y) ;
13 | **end**
14 **end**

 ISBN: 978-1-61208-162-5

Example 5: The results of the LHS-CHI2 algorithm with the relation example r (cf. Table I) using thresholds $MinSupCT = 0.2$, $MinPrec = 0.3$ and $MinCor = 1.8$ are given in Table III.

Table III
RESULT OF LHS-CHI2 ALGORITHM.

Correlation Rule	χ^2 value
BG	3.75
FG	4.44
BCF	4.24
BCG	9.10
BDF	10.14
CFG	5.74
DFG	4.93
BCFG	20.09

III. LITERAL CORRELATION RULES

In this section, we present our contribution. The aim is to build Literal Correlation Rules (Correlation Rules over Literal Patterns) only from (i) the relation r and (ii) the set of items $\mathcal{R}$. Mining such rules with the help of the relation $\overline{r}$ and of the set of items $\overline{\mathcal{R}}$ is not suitable because:

- Since the relation r is often sparse, the relation $\overline{r}$ is dense. As a result, the relational operators (union, intersection, ...) have slow performances over $r \cup \overline{r}$.
- It is possible to find an item $A \in \mathcal{R}$ such that the pattern $A\overline{A}$ satisfies all the constraints over a Correlation Rule. It is the case of the pattern B in our example relation. As a consequence, the set of solutions is polluted by inconsistent patterns.

We first define the concept of Literal Correlation Rule: an extension of Correlation Rules. We show, in a second step, that two literal patterns, which belong to the same Canonical Base, have the same χ^2 value and satisfy the same set of constraints (see Section II-B). As a consequence, the set of Correlation Rules can be used as a base for the Literal Correlation Rules. Then we show that the number of Literal Correlation Rules is exponential with regard to the number of Correlation Rules. We modify the LHS-CHI2 Algorithm in order to compute Literal Correlation Rules, and we add another pruning step in order to limit the number of results. We finally define the χ^2 function for a literal pattern $X\overline{W}$ as follows:

$$\chi^2(X\overline{W}) = \sum_{Y\overline{Z} \in \mathbb{P}(X\overline{W})} \frac{(Supp(Y\overline{Z}) - E(Y\overline{Z}))^2}{E(Y\overline{Z})} \quad (3)$$

Definition 4 (Literal (Decision) Correlation Rule): Let $X\overline{W}$ be a pattern, $MinCor$, $MinPerc$, $MinSupCT$ and $MaxCard$ thresholds specified by the end-user. According to the criteria introduced in Section II-B, the literal $X\overline{W}$ is a valid Literal Correlation Rule if and only if:

1) $\chi^2(X\overline{W}) \geq MinCor$;
2) $2 \leq |X\overline{W}| \leq MaxCard$;
3) $MinPerc$ cells of its contingency table have a support greater or equal than $MinSupCT$.

Moreover, if a value of the target attribute is present either in the positive part of the literal either in its negative part, the rule is called a Literal Decision Correlation Rule.

Example 6: Let us consider the following thresholds $MinCor = 1.8$, $MinSupCT = 0.2$ and $MinPrec = 0.3$, and the literal $B\overline{G}$. The contingency table associated to this literal pattern is:

	B	$\overline{B}$
G	2	2
$\overline{G}$	2	4

Since the four cells of this contingency table are greater than 2, we satisfy the third condition. We have $\chi^2(B\overline{G}) \simeq 3.75 \geq MinCor$ and the first condition is valid. Literal $B\overline{G}$ has a cardinality equal to 2, thus the second condition is checked. Moreover, Literal $B\overline{G}$ contains a value of the target attribute. As a consequence, the Literal Decision Correlation Rule materialized by $B\overline{G}$ is valid.

Let $X \cup \overline{A}$ and $X \cup A$ be two literal patterns, where X does not contain a negative part. The following lemma shows that the χ^2 values for both literal patterns are equals.

Lemma 1: Let X be a pattern and A a 1-item. We have: $\chi^2(X \cup \overline{A}) = \chi^2(X \cup A)$

The following proposition shows that any literal pattern belonging to the same Canonical Base has the same χ^2 value.

Proposition 1: Let $X \subseteq \mathcal{R}$ be a pattern, then we have:

$$\forall Y\overline{Z} \in \mathbb{P}(X), \chi^2(X) = \chi^2(Y\overline{Z}) \quad (4)$$

The following lemma indicates how we can build valid Literal Correlation Rules given only valid Correlation Rules.

Lemma 2: If a pattern X is a valid Correlation Rule (its χ^2 value is greater or equal than the threshold $MinCor$ and X satisfies all the constraints given in Section II-B), then any literal pattern belonging to its Canonical Base represents a valid Literal Correlation Rule.

Consequences of Proposition 1 and of Lemma 2 are very attractive. When mining Literal Correlation Rules, we do not need, as input of our algorithm, the set $\mathcal{R} \cup \overline{\mathcal{R}}$ but only the set $\mathcal{R}$. We just have to modify the processing done on the leaves of the LHS-CHI2 execution tree, in order to explore the Canonical Base associated with the current pattern. Moreover, as expected in introduction, we do not need the relation $\overline{r}$. Finally, the following results holds:

Corollary 1: Correlation Rules are a lossless representation for Literal Correlation Rules.

The concept of lossless representation applied to association rules [12] or to literal association rules mining [13], are very helpful to reduce the number of rules. However, we cannot predict an exact value for the expected gain. With the following lemma, we show that the number of Literal

Correlation Rules depends on the number of Correlation Rules having cardinality i ($i \in [2, MaxCard]$).

Lemma 3: Let us denote by *Sol* the set of solutions related to the problem of finding all the Correlation Rules satisfying all the constraints. Let Sol_i be the subset of *Sol* which contains only rules of cardinality i. Let *Sol'* be the set of solutions related to the problem of finding all the Literal Correlation Rules satisfying the same set of constraints than *Sol*. Then we have: $|Sol'| = \sum_{i=2}^{i=MaxCard} |Sol_i| * 2^i$.

A drawback highlighted by decision makers using the *MineCor* software (the software which implements the LHS-CHI2 Algorithm) is that the extracted rules which have a large χ^2 value could appear seldom in the relation. As a consequence, they consider that the obtained information is not of great quality. To answer their expectations, we modify the LHS-CHI2 Algorithm by adding a pruning step based on the support (using a threshold $MinSup$) and by extracting Literal Correlation Rules. The changes only affect the first three lines of the LHS-CHI2 Algorithm. The new algorithm is called LHS-LCHI2. Like in the LHS-CHI2 Algorithm, if we want to extract the Literal Correlation Rules and not only the Literal Decision Correlation Rules, we have to prune the test "$\exists t \in \mathcal{T} : t \in X$" for line 1 of the LHS-LCHI2 Algorithm.

Algorithm 2: LHS-LCHI2 Algorithm.

```
1 if Y = ∅ and ∃t ∈ T : t ∈ X and |X| ≥ 2 and
    χ²(X) ≥ MinCor then
2   foreach Y Z̄ ∈ ℙ(X) do
3     if Supp(Y Z̄) ≥ MinSup then
4       Output Y Z̄, χ²(X)
5     end
6   end
7 end
8 ...
```

Let us emphasize that the addition of the constraint "$Supp(Y\overline{Z}) \geq MinSup$" has the negative effect of making false Corollary 1 and Lemma 3 unless $MinSup$ equals 0.

Example 7: Continuing our example with parameters $MinSupCT = 0.2$, $MinPrec = 0.3$, $MinCor = 1.8$ and $MinSup = 0.4$, the results of the LHS-LCHI2 Algorithm are given in Table IV.

Table IV
RESULT OF LHS-CHI2 ALGORITHM.

Correlation Rule	χ^2 value	Support
$B\overline{G}$	3.75	6
FG	4.44	4
$B\overline{C}\,\overline{G}$	9.10	4
$BF\overline{D}$	10.14	4

IV. EXPERIMENTAL EVALUATIONS

Some representative results of the LHS-LCHI2 Algorithm are presented below. As emphasized in Section I, the experiments were done on different CSV files of real value measures supplied by STMicroelectronics (STM) and ATMEL (ATM). These files have one or more target columns, resulting from the concatenation of several measurement files. The characteristics of the relations used can be found in Table IV. All experiments were conducted on an HP Workstation (1.8 GHz processor with a 4 Gb RAM). To carry out pre-processing and transformation of these files into a binary relation, we implemented methods described in [14].

Table V
DATASET EXAMPLES

Name	Number of Columns	Number of Rows
STM File	1 281	297
ATM File	749	213

Figure 1. Number of Literal Decision Correlation Rules. Results with 4 intervals, $CtPerc = 0.34$, $MinCorr = 1.6$, $MinSupCT = 0.24$ (STM File - target1)

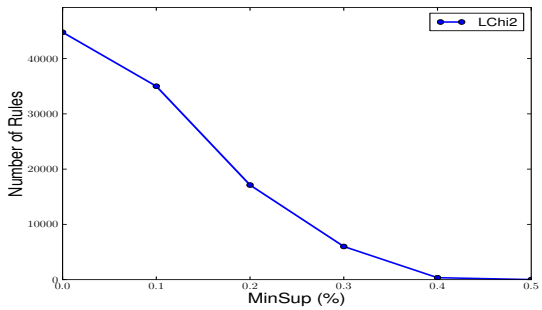
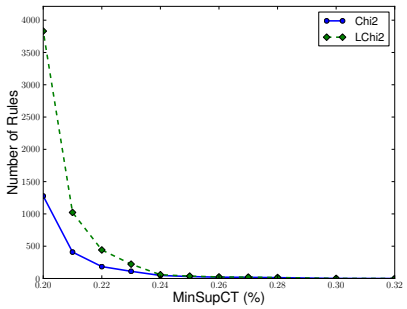


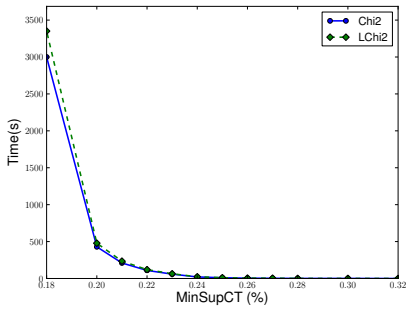
Figure 1 shows the impact of the $MinSup$ threshold over Literal Decision Correlation Rules. In the same way, when extracting frequent pattern, if the threshold $MinSup$ is large, no rule is produced. The lower $MinSup$ is, the more we can approach the bound given in lemma 3.

The goal of Figure 2(a) is to compare the number of rules produced by LHS-CHI2 and LHS-LCHI2 Algorithms. Since LHS-CHI2 Algorithm does not have a pruning step using the $MinSup$ threshold, we decided to fix it to the value of $MinSupCT$. The number of rules produced by the LHS-LCHI2 Algorithm is greater with a factor between 1 and 2.5. In Figure 2(b), we compare the two algorithms over the same hypothesis. As we can see, execution times are very close (less than 12% in the worst case). This can be explained by our specific implementation of the LHS-LCHI2 Algorithm: the computation of the χ^2 function and the pruning step using $MinSup$ both require the browsing of a contingency table. During the χ^2 computation, we

Figure 2. Results with 6 intervals, $CtPerc = 0.3$, $MinCorr = 2.8$, $MinSup = MinSupCT$ (ATM file - target3).



(a) Number of Decision Correlation Rules *vs.* Number of Literal Decision Correlation Rules



(b) Execution Time

put into a vector all the literal patterns having a support greater than $MinSup$. As a consequence, line 3 can be resumed as a browsing vector (which contains, in the worst case, $2^{MaxCard}$ elements). Thus, the difference between the execution times can be explained by the number of input/output operations which are more important in the LHS-LCHI2 Algorithm since we extract more rules.

V. Conclusion and Future Work

When mining Correlation Rules, one drawback is that the extracted rules which have a large χ^2 value appear seldom in the relation. As a consequence, we do not know which literal patterns have an important impact on the rules. To solve this problem, we have introduced the concept of Literal Correlation Rules: Correlation Rules admitting literal patterns. We show that the set of Correlation Rules satisfying a set of constraints is a base for the Literal Correlation Rules satisfying the same set of constraints. Thus we provide an upper border for the number of Literal Correlation Rules. In order to highlight only relevant Literal Correlation Rules, we add a pruning step based on the support of a literal, and therefore modified the related algorithm.

To continue our work, we intend to use multi-core strategies. In a first step, one thread could process the leaves of the execution tree while another could explore the branches of the tree. In a second step, our aim is to to parallelize each branch of the LHS-LCHI2 algorithm.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996, pp. 307–328.

[2] G. Chen, H. Liu, L. Yu, Q. Wei, and X. Zhang, "A new approach to classification based on association rule mining," *Decision Support Systems*, vol. 42, no. 2, pp. 674–689, 2006.

[3] W. Li, J. Han, and J. Pei, "Cmar: Accurate and efficient classification based on multiple class-association rules," in *ICDM*, IEEE Computer Society, 2001, pp. 369–376.

[4] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules," *ACM Trans. Inf. Syst.*, vol. 22, no. 3, pp. 381–405, 2004.

[5] S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," in *SIGMOD Conference*, 1997, pp. 265–276.

[6] A. Casali and C. Ernst, "Extracting decision correlation rules," in *DEXA*, ser. Lecture Notes in Computer Science, vol. 5690. Springer, 2009, pp. 689–703.

[7] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

[8] M. Spiegel and L. Stephens, *Outline of Statistics*. McGraw-Hill, 1998.

[9] D. Moore, "Measures of lack of fit from tests of chi-squared type," in *Journal of statistical planning and inference*, vol. 10, no. 2, 1984, pp. 151–166.

[10] M. Laporte, N. Novelli, R. Cicchetti, and L. Lakhal, "Computing full and iceberg datacubes using partitions," in *ISMIS*, 2002, pp. 244–254.

[11] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 241–258, 1997.

[12] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal, "Generating a condensed representation for association rules," *J. Intell. Inf. Syst.*, vol. 24, no. 1, pp. 29–60, 2005.

[13] G. Gasmi, S. B. Yahia, E. M. Nguifo, and S. Bouker, "Extraction of association rules based on literalsets," in *DaWaK*, ser. Lecture Notes in Computer Science, vol. 4654. Springer, 2007, pp. 293–302.

[14] D. Pyle, *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

 ISBN: 978-1-61208-162-5

Designing Cost-sensitive Fuzzy Classification Systems Using Rule-weight

Mansoor Zolghadri Jahromi[1], Mohammad Reza Moosavi[2]
School of Electrical and Computer Engineering, Shiraz University,
Shiraz, Iran
[1]zjahromi@shirazu.ac.ir, [2]mmoosavi@cse.shirazu.ac.ir

***Abstract*— In the field of pattern classification, we often encounter problems that class-to-class misclassification costs are not the same. For example, in the medical domain, misclassifying a patient as normal is often much more costly than misclassifying a normal as patient. Our aim in this paper is to propose a method of designing fuzzy rule-based classification systems to tackle this problem. We use rule-weight as a simple mechanism to tune the rule-base. Assuming that class-to-class misclassification costs are known, we propose a learning algorithm that attempts to minimize the total cost of the classifier on train data (i.e., instead of minimizing the error-rate). Using a number of UCI datasets we show that the method is quite effective in reducing the average cost of the classifier on test data.**

Keywords- Fuzzy Classification Systems; Cost Sensitive Classification; Rule Weight; Data Mining

I. INTRODUCTION

A Fuzzy Rule-Based Classification System (FRBCS) is a special case of fuzzy modeling where the output of the system is crisp and discrete. Basically, the design of a FRBCS consists of finding a compact set of fuzzy if-then classification rules to be able to model the input-output behavior of the system. The information available about the behavior of the system is assumed to be a set of input-output example pairs (i.e., a number of pre-labeled classification examples).

The most challenging problem in designing FRBCSs is the construction of rule-base for a specific problem. Many approaches have been proposed to construct the rule-base from numerical data. These include heuristic approaches [1, 2], neuro-fuzzy techniques [3-5], clustering methods [6-8], genetic algorithms [9-12] and data mining techniques [13-15].

One main advantage of fuzzy rule-based systems in classification problems is their interpretability. Using linguistic labels in the antecedent of the fuzzy rules makes them very understandable, which is the main characteristic of this type of classifier.

There are many classification problems where class-to-class misclassification costs are different. For example, in medical diagnosis of cancer, diagnosing malignant tumors as benign and hence treating a cancer patient as healthy could be much more costly than interpreting benign tumors as malignant.

Cost-sensitive learning first introduced by Elkan [16] has shown to be an effective technique for incorporating the different misclassification costs into the classification process [17-21].

A pattern classification problem can be easily reformulated as a cost minimization problem. In [22], the concept of instance weight is introduced for each training pattern in order to handle the cost-sensitive problems. The weight of an input pattern represents the average cost of misclassifying that pattern. Fuzzy if–then rules are generated by considering the weights as well as the compatibility of training patterns. A rule-weight learning method based on Reward and Punishment is also proposed to tune the weight of the rules.

In this paper, we assume that for the problem in hand a cost matrix C giving class-to-class misclassification costs is available. We assume that the cost of misclassifying an instance depends on its actual and predicted classes. Each element c_{ij} of this matrix gives the cost of classifying a pattern from class i in class j (c_{ij}=0 if i=j). This is slightly different from the scheme that assumes that the misclassification cost of an instance depends only on its actual class [16, 20]. The design of the classifier is then viewed as a cost minimization problem.

For a specific cost-sensitive problem, an initial rule-base is constructed using one of the methods proposed in the literature [22]. The initial rule-base is then tuned to the problem in hand by assigning a weight to each fuzzy rule in the constructed rule-base. The novelty of our method is in the rule-weight learning algorithm that we propose. The proposed algorithm uses the cost matrix to directly minimize the total misclassification cost of the classifier on training data. In this process, the size of the rule-base is reduced by assigning zero weight to redundant rules, which improves the interpretability of the final rule-base. Using a number of datasets from UCI ML repository, we show that the scheme is quite effective in constructing a compact rule-base for cost-sensitive problems.

The rest of this paper is organized as follows. In Section II, the structure of a fuzzy classification system is introduced. In Section III, a method of constructing the rule-base for conventional (i.e., not cost sensitive) problems is discussed. In Section IV, a method of constructing rule-base for cost-sensitive problems is presented. In Section V, the proposed method of rule-weight learning is presented. In Section VI, the simulation results are presented. Section VII concludes this paper.

 ISBN: 978-1-61208-162-5

II. FUZZY RULE-BASED CLASSIFICATION SYSTEMS

A fuzzy rule-based classification system is composed of three main conceptual components: database, rule-base, and reasoning method. The database describes the semantic of fuzzy sets associated to linguistic labels. Each rule in the rule-base specifies a subspace of pattern space using the fuzzy sets in the antecedent part of the rule. The reasoning method provides a mechanism to classify a pattern using the information from the rule-base and database.

Different rule types have been used for pattern classification problems [23]. We use fuzzy rules of the following type for an n-dimensional problem:

$$\text{Rule } R_j\text{: If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \text{ then class } h \text{ with } CF_j, \quad j=1, 2, \ldots, N \tag{1}$$

where $X=[x_1, x_2, \ldots, x_n]$ is the input feature vector, $h \in \{C_1, C_2 \ldots, C_M\}$ is the label of the consequent class, A_{jk} is the fuzzy set associated to x_k , CF_j is the certainty grade (i.e., rule weight) of rule R_j and N is the number of fuzzy rules in the rule-base.

In order to classify an input query pattern $X_t = [x_{t1}, x_{t2}, \ldots, x_{tn}]$, the degree of compatibility of the pattern with each rule is calculated (i.e., using a T-norm to model the "and" connectives in the rule antecedent). In case of using product as T-norm, the compatibility grade of rule R_j with the input pattern X_t can be calculated as:

$$\mu_j(X_t) = \prod_{i=1}^{n} \mu_{A_{ji}}(x_{ti}) \tag{2}$$

Using single winner reasoning method, an input query pattern X_t is classified according to the consequent class of the winner rule R_w. With the rules of form (1), the winner rule R_w is identified as:

$$w = \arg\max_{1 \le j \le N} \{\mu_j(X_t).CF_j\} \tag{3}$$

III. RULE-BASE CONSTRUCTION

For an M-class problem in an n-dimensional feature space, assume that m labeled patterns of the form $X_p=[x_{p1}, x_{p2}, \ldots, x_{pn}]$, $p=1, 2, \ldots, m$ are given. A simple approach for generating fuzzy rules is to partition the domain interval of each input attribute using a pre-specified number of fuzzy sets (i.e., grid partitioning). Some examples of this partitioning (using triangular membership functions) are shown in Fig. 1.

Given a partitioning of pattern space, one approach is to consider all possible combination of the antecedents to generate the fuzzy rules. The selection of the consequent class for an antecedent combination (i.e., a fuzzy rule) can be easily expressed in terms of confidence of an association rule from the field of data mining [24]. A fuzzy classification rule can be viewed as an association rule of the form $A_j \Rightarrow class\ C_j$ where, A_j is a multi-dimensional fuzzy set representing the antecedent conditions and C_j is a class label. Confidence of a fuzzy association rule R_j is defined as [15]:

$$C(A_j \Rightarrow class\ C_j) = \frac{\sum_{X_p \in class\ C_j} \mu_j(X_p)}{\sum_{p=1}^{m} \mu_j(X_p)} \tag{4}$$

where $\mu_j(X_p)$ is the compatibility grade of pattern X_p with the antecedent of the rule R_j, m is the number of training patterns, and C_j is a class label.

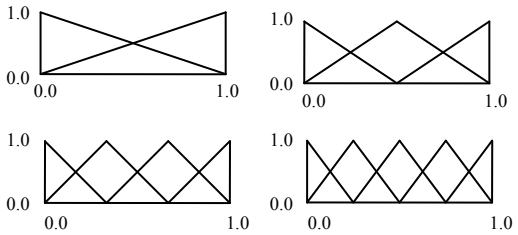


Figure 1. Different partitioning of each feature axis.

A common approach for identifying the consequent class C_q of an antecedent combination A_j is to specify the class with maximum confidence as the consequent class. This can be expressed as:

$$q = \arg\max_{1 \le h \le M} \left\{C(A_j \Rightarrow Class\ C_h)\right\} \tag{5}$$

The problem with grid partitioning is that an appropriate partitioning of each attribute is not usually known. One solution for this is to simultaneously consider different partitioning of an attribute (see Fig. 1). That is, for each attribute, a pre-specified number of fuzzy sets (for example 14, as shown in Fig. 1) can be used when generating a fuzzy rule. The problem is that for an n-dimensional problem, 14^n antecedent combinations should be considered. It is impractical to consider such a huge number of antecedent combinations when dealing with high dimensional problems.

One solution for the above problem is presented in [15] by adding the fuzzy set "don't care" to each attribute. The membership function of this fuzzy set is defined as $\mu_{don't\ care}(x) = 1$ for all values of x. The trick is not to consider all antecedent combinations (which is now 15^n) and only short fuzzy rules having a limited number of antecedent conditions (excluding don't care) are generated as candidate rules.

The number of candidate rules generated with the above scheme can still be quite large for many problems. A compact rule-base can be constructed in the following manner. The generated candidate rules are divided into M groups according to their consequent classes. The candidate rules in each group are sorted in descending order of an evaluation criterion. A rule-base is constructed by choosing Q fuzzy rules from each class (i.e., $M \times Q$ fuzzy rules in total). Among many heuristic rule evaluation measures presented in the literature [25], we use the measure presented in [10]. The

 ISBN: 978-1-61208-162-5

evaluation of rule R_j (i.e., $A_j \Rightarrow class\ C_j$) with this measure can be expressed as:

$$e(R_j) = \sum_{X_p \in Class\ C_j} \mu_j(X_p) - \sum_{X_p \notin Class\ C_j} \mu_j(X_p) \quad (6)$$

IV. COST-SENSITIVE FUZZY CLASSIFICATION SYSTEMS

For an M-class problem, assume that an $M \times M$ cost matrix C giving class-to-class misclassification costs is given. In this section, we extend the rule-base construction method of the previous section for the case of cost-sensitive problems. For this purpose, we assign a weight to each training example. The weight assigned to each training example is the average cost of classifying that example. The cost matrix C can be used to calculate the weight w_p a training example X_p (from class i) as:

$$w_p = \frac{1}{M} \sum_{j=1}^{M} c_{i,j} \quad (7)$$

where, $c_{i,j}$ denotes the cost of classifying an instance of class i in class j. The weight assigned to a training example can be viewed as the importance of that pattern in the classification process. Using the weights of the training examples, the confidence of an association rule (4) can be easily modified to:

$$C(A_j \Rightarrow class\ C_j) = \frac{\sum_{X_p \in class\ C_j} w_p . \mu_j(X_p)}{\sum_{p=1}^{m} w_p . \mu_j(X_p)} \quad (8)$$

The rule evaluation metric (6) is modified to accommodate the weights assigned to training examples:

$$e(R_j) = \sum_{X_p \in Class\ C_j} w_p . \mu_j(X_p) - \sum_{X_p \notin Class\ C_j} w_p . \mu_j(X_p) \quad (9)$$

It must be noted that equations (8) and (9) cover the special case of cost-insensitive problems (i.e., w_p=1, p=1,2,..,m). In short, the rule generation process discussed in Section III can be used to construct a rule-base for a cost-sensitive problem. For this purpose, equations (4) and (6) are replaced by (8) and (9), respectively.

V. RULE-WEIGHT LEARNING ALGORITHM

For the problem in hand, assume that a rule-base consisting of N fuzzy classification rules $\{R_j, j=1, 2, ..., N\}$ is constructed using the method discussed in the previous section. Our aim in this section is to propose a rule-weight learning algorithm that attempts to minimize the total cost misclassification cost of the rule-base on train data. For this purpose, we make use of the rule-weight learning algorithm that was proposed in [26], which attempts to minimize the error-rate of the classifier on training data. In this section, we propose an extended version of this algorithm to cover case-sensitive problems. The proposed algorithm attempts to minimize the total misclassification cost of the constructed rule-base on the training data.

In its basic form, the proposed algorithm is a hill-climbing search method. The algorithm starts with an initial solution to the problem (i.e., $\{CF_k = 1, k = 1, 2, \ldots, N\}$) and attempts to improve the solution by adjusting the weight of each rule in turn (to reduce the total cost on train data). The basic component of the learning scheme is an algorithm (denoted as *best-weight*) that provides the answer to the following question: *"What is the optimal weight of a rule (i.e., R_k) assuming that the weights of all other rules are given and fixed?"*

The weight found by *best-weight* is optimal in the sense that it results in minimum total misclassification cost on training data. In this way, the overall learning algorithm consists of visiting each rule in turn to adjust its weight. It must be noted that the weight specified for a rule is optimal if the weights of other rules in the rule-base remain fixed. That is why the second pass and subsequent passes over the rules can reduce the cost on train data. In experiments, as a mechanism to prevent overfitting, we stop the search after a fixed number of passes over all rules [26].

To illustrate how the *best-weight* algorithm finds the optimal weight of a rule, consider rule R_k for optimization.

$$\text{Rule } R_k\text{: If } x_1 \text{ is } A_{k1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{kn} \text{ then class } T \text{ with } CF_k \quad (10)$$

To calculate the optimal weight of rule R_k (i.e., CF_k), the rule is first removed from the rule-base by setting its weight to zero (CF_k=0). In the next step, the predicted class of all the training patterns will be found and stored (without rule R_k in the rule-base). Then, the score S of each training data X_t in covering subspace of rule R_k (i.e., $\mu_k(X_t) \neq 0$) is calculated using the following definition of score:

$$S(X_t) = \frac{\max_{1 \le j \le N} \{CF_j . \mu_j(X_t) \mid R_j \neq R_k\}}{\mu_k(X_t)} \quad (11)$$

where, $\mu_k(X_t)$ denotes the compatibility grade of pattern X_t with rule R_k . For a pattern X_t having score $S(X_t)=a$, if we choose $CF_k > a$, the pattern X_t will be classified by rule R_k (i.e., as class T) since rule R_k will be the winner rule (having maximum weighted compatibility with pattern X_t). In case we choose $CF_k < a$, the pattern will be classified as if we don't have rule R_k in the rule-base (we have already stored the predicted class in previous step).

For a specific value of CF_k, the predicted class of X_t (with $S(X_t)=a$) for the two interval of $CF_k < a$ and $CF_k > a$ are known. As we know the true class of X_t, we can easily calculate the cost of classifying X_t for $CF_k < a$ and $CF_k > a$. For a training pattern X_t, assume that L is the true class, P is the predicted class for $CF_k < a$, and T is the predicted class for $CF_k > a$. Then, the cost of classifying X_t for $CF_k < a$ and $CF_k > a$ can be expressed as:

$$Cost(X_t) = \begin{cases} C_{T,P} & if\ CF_k < a \\ C_{T,L} & if\ CF_k > a \end{cases} \quad (12)$$

where, $C_{I,J}$ is used to represent the cost of classifying an instance of class I in class J.

Having the relation between a certain value of CF_k and the corresponding total cost of training data, the best value of CF_k can be easily found by sorting the patterns in ascending order of their scores (i.e., $S(X_1) < S(X_2) < .... < S(X_n)$). Considering any value of CF_k between $S(X_i)$ and $S(X_{i+1})$, the first i patterns will be classified as class T and the rest of the patterns will be classified as if rule R_k is not in rule-base. In this way, $n+1$ different values of CF_k should be examined to find its best value. The *best-weight* algorithm for calculating the best weight of a rule is given in Fig. 2.

The algorithm starts by finding the predicted class of each pattern when the rule is removed from the rule-base. The patterns are then sorted in ascending order of their scores. For a rule having n training pattern in its covering space, the algorithm of Fig. 2 examines $n+1$ values to find the best weight (*best-CF*) for the rule. The first and last values are "zero" and "*last*+ε", respectively (last is the score of last pattern in the ranked list and ε is a very small positive number). The rest are examined in the middle of two successive scores.

VI. EXPERIMENTAL RESULTS

In order to assess the performance of the proposed method, we used four data sets available from UCI ML repository. Some statistics of these datasets are shown in Table I.

To construct an initial rule-base for a specific problem, we used the method of Section III to generate rules of *length*≤2. The candidate rules were the grouped based on their consequent classes. The rules in each group were then sorted according to the rule evaluation metric. An initial rule-base was constructed by choosing a certain number of best rules from each group. The proposed rule-weight learning algorithm was used to optimize the rule-base by passing 4 iterations over all rules.

We used 10-times 10-fold cross validation technique to assess the generalization ability of the proposed method. In each fold, 90% of the data were use to construct the rule-base. The proposed rule-weight learning algorithm was then used to specify the weights of all rules in the rule-base. The performance on test data was measured by calculating the average cost per example (CPE).

For the purpose of experiments we assumed a cost matrix using the number of instances in each class (i.e., class proportionate cost). The misclassification cost of predicting an instance of class i in class j is assumed to be:

$$C_{ij} = \frac{\text{(total no. of patterns of class j)}}{\text{(total no. of patterns of class i)}} \quad (13)$$

This cost matrix assumes that misclassification of the minority class (with a small number of training patterns) is more costly than majority class. In Tables II and III we give the cost matrix used for each dataset, which is based on equation (13).

It must be noted that the misclassification costs in most medical problems depends strongly on the domain, and particularly the anticipated consequences of the misclassification. This is not directly related to the class proportions. These cost matrices are used as examples (i.e., they don't represent the actual misclassification costs) to evaluate the proposed rule-weight learning algorithm.

```
Inputs: training patterns in the covering subspace of the rule and true class of each
pattern {(Xt, true-class(Xt)), t=1,2,…,n}

Output: the best weight for the rule (best-CF) assuming that the weights of all other
rules are fixed

CF = 0    (i.e. remove the rule from the rule-base)
for each training pattern, Xi
      Calculate and memorize the predicted class of Xi
      Calculate and memorize S(Xi) using eq. 11
rank the patterns in ascending order of their scores in a list

#assume that Xk and Xk+1 are two successive patterns in the list
#also assume that Xlast is the last pattern in the list and ε is a small positive #number

for each value of CF (i.e. CF=0, CF=(Score(Xk)+Score(Xk+1))/2, CF=Score(X last)+ε)
      Calculate and memorize total misclassification cost corresponding to the specified
      value of CF

best_CF = CF with minimum total misclassification cost

return best-CF
```

Figure 2. Best-Weight Algorithm for finding the best weight of a rule

In Table IV, we report the CPE for the initial rule-base (i.e., before rule-weighting) and after applying the rule-weighting algorithm of Section IV. As seen, our rule-weighting algorithm has significantly reduced the CPE on test data for all datasets used in our experiments.

In order to assess the performance of our method in comparison with other methods proposed in the literature, in Table V, we report the results of the method proposed in [22] to handle cost sensitive problems. The cost matrixes used to produce the results of this Table is the same as Table IV (i.e., class proportionate cost matrix (13)).

Comparing the results of Tables IV and V, we observe that our proposed rule-weighting algorithm outperforms the method proposed in [22] by achieving lower value of CPE on test data for all datasets used in experiments, which was the primary goal of the algorithm.

TABLE I. SOME STATISTICS OF THE DATASETS USED IN EXPERIMENTS.

Dataset	No. Of features	No. of instances	No. of classes	No. of instances per class
Thyroid	5	215	3	35, 30, 150
Pima	8	768	2	500, 268
Bupa	6	345	2	145, 200
Breast cancer	30	569	2	357, 212

TABLE II. CLASS-PROPORTIONAL COST MATRIX FOR THYROID DATASET.

	hyper-thyroidism	hypo-thyroidism	normal
hyperthyroidism	0	0.86	4.28
hypothyroidism	1.17	0	5
normal	0.23	0.2	0

TABLE III. CLASS-PROPORTIONAL COST MATRICES.

Pima	***tested_negative***	***tested_positive***
tested_negative	0	0.536
tested_positive	1.87	0
Bupa	***drinks<5***	***drinks>5***
drinks<5	0	1.38
drinks>5	0.73	0
Breast cancer	***Benign***	***malignant***
benign	0	0.594
malignant	1.68	0

TABLE IV. THE CPE ON TRAIN AND TEST DATA FOR VARIOUS DATASETS USING OUR PROPOSED METHOD.

Dataset	Train data		Test data	
	Before rule-weighting	After rule-weighting	Before rule-weighting	After rule-weighting
Thyroid	1.24	0.03	1.31	0.12
Pima	1.42	0.42	1.43	0.54
Bupa	0.58	0.26	0.59	0.36
Breast cancer	0.12	0.03	0.13	0.06

TABLE V. THE CPE ON TRAIN AND TEST DATA FOR VARIOUS DATASETS USING THE METHOD PROPOSED IN [22].

Dataset	Train data		Test data	
	Before rule-weighting	After rule-weighting	Before rule-weighting	After rule-weighting
Thyroid	0.25	0.15	0.25	0.2
Pima	0.96	0.79	0.97	0.8
Bupa	0.57	0.4	0.58	0.42
Breast cancer	0.52	0.24	0.55	0.24

In Table VI, we report on average number of rules in the final rule-base using our method. As seen, the number of rules in the final rule-base is much smaller than initial rule-base. This is due to the fact that our algorithm removes the redundant rules by setting their weights to zero. This is important since the interpretability and efficiency of the rule-base is improved.

TABLE VI. AVERAGE NUMBER OF RULES IN THE FINAL RULE-BASE USING THE PROPOSED METHOD.

Dataset	Before rule-weighting	After rule-weighting
Thyroid	99	4.53
Pima	66	10.5
Bupa	66	11.9667
Breast cancer	50	6.2

VII. CONCLUSIONS

In this paper, a cost-sensitive learning algorithm was proposed to tune a fuzzy classification system by specifying the weights of fuzzy rules. The learning algorithm makes use of the cost matrix giving class-to-class misclassification costs to minimize the total cost on train data.

 ISBN: 978-1-61208-162-5

Using a number of real-life datasets, we showed that the scheme is quite effective in reducing the average cost of the classifier on test data. Another advantage of the proposed method is that redundant rules are removed during the learning process. This feature is very useful since the final rule-base is better in terms of interpretability and classification speed.

Since the proposed learning method attempts to minimize the classification cost of the classifier on training data, obviously, this can cause the classifier to overfit the training data. The main cause for this is that the learning algorithm does not have a mechanism to cope with noisy training examples (i.e., those in contradiction with the rest of training patterns). A mechanism is needed to deal with this issue.

REFERENCES

[1] S. Abe and M. S. Lan, "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification," IEEE Trans. on Fuzzy Systems, vol. 3, Feb. 1995, pp. 18-28.

[2] H. Ishibuchi, K. Nozaki and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," Fuzzy Sets and Systems, vol. 52, Nov. 1992, pp. 21-32 .

[3] S. Mitra and L. I. Kuncheva, "Improving classification performance using fuzzy MLP and two-level selective partitioning of the feature space," Fuzzy Sets and Systems, vol. 70, Feb. 1995, pp. 1-13.

[4] D. Nauck and R. Kruse, "A neuro-fuzzy method to learn fuzzy classification rules from data," Fuzzy Sets and Systems, vol. 89, Aug. 1997, pp. 277-288.

[5] I. Gadaras and L. Mikhailova, "An interpretable fuzzy rule-based classification methodology for medical diagnosis," Artificial Intelligence in Medicine, vol. 47, Sep. 2009, pp. 25-41.

[6] J. A. Roubos, M. Setnes, and J. Abonyi, "Learning fuzzy classification rules from labeled data," Information Sciences, vol. 150, Mar. 2003, pp. 77-93.

[7] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," Pattern Recognition Letters, vol. 24, Feb. 2003, pp. 2195–2207.

[8] P. Pulkkinen and H. Koivisto, "Identification of interpretable and accurate fuzzy classifiers and function estimators with hybrid methods," Applied Soft Computing, vol. 7, Mar. 2007, pp. 520-533.

[9] J. Casillas, O. Cordón, and M. J. Del Jesus, F. Herrera, "Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems," Information Sciences, vol. 136, Aug. 2001, pp. 135-157.

[10] A. Gonzalez and R. Perez, "SLAVE: A genetic learning system based on an iterative approach," IEEE Trans. on Fuzzy Systems, vol. 7, Apr. 1999, pp. 176-191.

[11] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," Information Sciences, vol. 136, Aug. 2001, pp. 109-133.

[12] L. Sánchez, I. Couso, J. A. Corrales, O. Cordón, M. J. Del Jesus, and F. Herrera, "Combining GP operators with SA search to evolve fuzzy rule based classifiers," Information Sciences, vol. 136, Aug. 2001, pp. 175-191.

[13] Y. Chung Hu and G. Hshiung Tzeng, "Elicitation of classification rules by fuzzy data mining," Engineering Applications of Artificial Intelligence, vol. 16, Oct. 2003, pp. 709-716.

[14] M. De Cock, C. Cornelis, and E. E. Kerre, "Elicitation of fuzzy association rules from positive and negative examples," Fuzzy Sets and Systems, vol. 149, Jan. 2005, pp. 73-85.

[15] H. Ishibuchi and T. Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining," Fuzzy Sets and Systems, vol. 141, Jan. 2004, pp. 59-88.

[16] C. Elkan, "The foundations of cost-sensitive learning," Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 01), June 2001, pp. 973-978.

[17] T. Nakashima, Y. Yokota, H. Ishibuchi, and G. Schaefer, "A cost-based fuzzy system for pattern classification with class importance," Artificial Life and Robotics, vol. 12, Sep. 2008, pp. 43-46.

[18] J. Zheng, "Cost-sensitive boosting neural networks for software defect prediction," Expert Systems with Applications, vol. 37, June 2010, pp. 4537–4543.

[19] S. Viaene and G. Dedene, "Cost-sensitive learning and decision making revisited," European Journal of Operation Research, vol. 166, Oct. 2005, pp. 212–220.

[20] L. Li, M. Chen, H. Wang, and H. Li, "CoSFuC: A Cost Sensitive Fuzzy Clustering Approach for Medical Prediction," Proc. of Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 08), pp. 127–131.

[21] G. Schaefer and T. Nakashima, "Application of Cost-sensitive Fuzzy Classifiers to Image Understanding Problems," Proc. Of IEEE International conference on Fuzzy Systems (FUZZ-IEEE 2009), Oct. 2009, pp. 1364–1368.

[22] T. Nakashima, G. Schaefer, Y. Yokota, and H. Ishibuchi, "A weighted fuzzy classifier and its application to image processing tasks," Fuzzy Sets and Systems, vol. 158, Feb. 2007, pp. 284-294.

[23] O. Cordon, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," International Journal of Approximate Reasoning, vol. 20, Jan. 1999, pp. 21-45.

[24] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 94), Dec. 1994, pp. 487-499.

[25] H. Ishibuchi and T. Yamamoto, "Comparison of heuristic criteria for fuzzy rule selection in classification problems," Fuzzy Optimization and Decision Making, vol. 3, June 2004, pp. 119-139.

[26] M. Zolghadri Jahromi and M. Taheri, "A Proposed Method for learning rule weights in Fuzzy Rule Based Classification Systems," Fuzzy Sets and Systems, vol. 159, Feb, 2008. pp. 449-459

 ISBN: 978-1-61208-162-5