DATA ANALYTICS 2012

The First International Conference on Data Analytics

ISBN: 978-1-61208-242-4

September 23-28, 2012

Barcelona, Spain

DATA ANALYTICS 2012 Editors

Sandjai Bhulai, VU University Amsterdam, The Netherlands

Joseph Zernik, Human Rights Alert (NGO), USA

Petre Dini, Concordia University, Canada / China Space Agency Center, China

DATA ANALYTICS 2012

Forward

The First International Conference on Data Analytics (DATA ANALYTICS 2012), held on September 23-28, 2012 in Barcelona, Spain, was an inaugural event on fundamentals in supporting data analytics, special mechanisms and features of applying principles of data analytics, application-oriented analytics, and target-area analytics.

Processing of terabytes to petabytes of data, or incorporating non-structural data and multi-structured data sources and types require advanced analytics and data science mechanisms for both raw and partially-processed information. Despite considerable advancements on high performance, large storage, and high computation power, there are challenges in identifying, clustering, classifying, and interpreting of a large spectrum of information.

We take here the opportunity to warmly thank all the members of the DATA ANALYTICS 2012 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 that dedicated much of their time and efforts to contribute to the DATA ANALYTICS 2012. We truly believe that thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the DATA ANALYTICS 2012 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success. We gratefully appreciate to the technical program committee co-chairs that contributed to identify the appropriate groups to submit contributions.

We hope the DATA ANALYTICS 2012 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in data analytics.

We hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

DATA ANALYTICS 2012 Chairs:
Sandjai Bhulai, VU University Amsterdam, The Netherlands
Joseph Zernik, Human Rights Alert (NGO), USA
Petre Dini, Concordia University, Canada / China Space Agency Center, China

DATA ANALYTICS 2012

Committee

DATA ANALYTICS 2012 Technical Program Committee

Rajeev Agrawal, North Carolina A&T State University - Greensboro, USA
Fabricio Alves Barbosa da Silva, Brazilian Army Technological Center - Rio de Janeiro, Brazil
Giuliano Armano, University of Cagliari, Italy
Ryan G. Benton, University of Louisiana at Lafayette, USA
Michael Berthold, Universität Konstanz, Germany
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Huiping Cao, New Mexico State University, USA
Michelangelo Ceci, University of Bari, Italy
Federica Cena, Università degli Studi di Torino, Italy
Qiming Chen, HP Labs - Palo Alto, USA
Been-Chian Chien, National University of Tainan, Taiwan
David Chiu, Washington State University, USA
Alain Crolotte, Teradata Corporation - El Segundo, USA
Tran Khanh Dang, National University of Ho Chi Minh City, Vietnam
Jérôme Darmont, Université de Lyon - Bron, France
Ernesto William De Luca, University of Applied Sciences Potsdam, Gernmany
Kamil Dimililer, Near East University, Cyprus
Shifei Ding, China University of Mining and Technology - Xuzhou City, China
Sourav Dutta, IBM Research Lab. - New Delhi, India
Sherif Elfayoumy, University of North Florida, USA
Yi Fang, Purdue University - West Lafayette, USA
Matjaz Gams, Jozef Stefan Institute - Ljubljana, Slovenia
Shlomo Geva, Queensland University of Technology - Brisbane, Australia
Ahmad Ghazal, Teradata Corporation - El Segundo, USA
Raju Gottumukkala, University of Louisiana at Lafayette, USA
William Grosky, University of Michigan - Dearborn, USA
Tudor Groza, The University of Queensland, Australia
Jerzy W. Grzymala-Busse, University of Kansas - Lawrence, USA
Shengbo Guo, Xerox Research Centre Europe, France
Tiziana Guzzo, National Research Council/Institute of Research on Population and Social Policies - Rome, Italy
Sven Hartmann, TU-Clausthal, Germany
Yi Hu, Northern Kentucky University - Highland Heights, USA
Jun (Luke) Huan, University of Kansas - Lawrence, USA
Sergio Ilarri, University of Zaragoza, Spain
Prabhanjan Kambadur, IBM TJ Watson Research Center, USA
Michal Kratky, VŠB-Technical University of Ostrava, Czech Republic
Dominique Laurent, University of Cergy Pontoise, France
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany

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

Complexity Analysis in the Language of Information Technology Companies

Mary Luz Mouronte López
Departamento de Ingeniería Telemática,
Universidad Carlos III de Madrid, Escuela Politécnica Superior
Av. Universidad 30, Edif. Torres Quevedo, Madrid 28911, Spain
mmouront@it.uc3m.es

***Abstract*—This paper uses a large amount of data, which, in turn, will interpret a broad spectrum of information. It applies the graph theory to map relationships among words using *network science* in order to analyze the reports of important telecommunication companies and to study the syntactic structure of their language. The following properties are analyzed in the word network of these companies, such as the words' co-occurrence, density, betweenness, distance between words (length, average diameter) and structures (presence of clusters, existence of motifs). We conclude that the company language shows characteristics of complex systems and that some features are common among organizations.**

Keywords-words network; mean distance; communities; motif; robustness.

I. INTRODUCTION

A language is a set of signs that allow human communication by means of their meaning and their relationships. It is different in creativity, form and content for several cultural groups because they do not use the same sounds, or grammatical structure [1]. Many tools for natural language processing have been developed due to language analysis, which is useful in different areas;

- The Internet has created many electronic documents, which are widely available. In order to recover the most relevant data, it is necessary to check these documents by means of language processing, a keyword-match based on data recovery, which supplies quick access to documents containing important information.
- In social marketing, understanding what the community is saying is a key issue. A study about abbreviations or variants of expression usage (small or capital letters, letter repetition or missing letters, orthographical variations, or non-alpha symbols) would be interesting.
- In medicine, the linguistic analysis may be useful in the investigation of mental illnesses. For instance, some studies have confirmed that schizophrenics have a syntactically less complex speech, which means that linguistic variable usage in a discriminant function analysis may help to predict diagnoses in many cases.

The objective of this paper is to study the syntactical structure of the language in information technology companies by means of *network science*. In [5], network science is defined as "the study of networks which contrasts, compares, and integrates techniques and algorithms developed in disciplines as diverse as mathematics, statistics, physics, social network analysis, information science and computer science".

This investigation studies more than 30 company reports with more than 6,000,000 words to interpret a broad spectrum of information. Specifically, the annual reports (in Spanish) of five companies are analyzed in: AMPER [23], JAZZTEL [24], VODAFONE [25], INDRA [26] and TELEFÓNICA [27].

We transform each company's annual report into a words network where some metrics such as: shortest distance between nodes, betweenness, detection of clusters and motifs are studied. These measures have also been used to analyze the structural properties of others natural and artificial complex systems [2] [3] [4]. Typical language properties are identified and the obtained results are compared among the analyzed companies.

We apply the method described in [6] to detect motifs and the Walktrap Algorithm [7] [22] to calculate clusters in the networks.

We develop several tools in C and phyton languages to analyze the language structure syntactically in the annual reports.

The rest of the paper is organized as follows: Section 2 shows the related work, Section 3 describes the method of analysis and results, and Section 4 establishes the main conclusions and the future work.

II. RELATED WORK

There are no previous works on applications about communication pattern analysis of information technology companies. Nevertheless, there are several works about the language characteristics.

Ferrer i Cancho and Sole [8] show that language organization, described in terms of a graph, displays two important features found in a disparate number of complex systems: (i) The so-called small-world effect. In particular, the average distance between two words, $< l >$, is shown

 ISBN: 978-1-61208-242-4

as $< l > \simeq 2^3$. (ii) A scale-free distribution of degrees and the fact that disconnecting the most connected nodes in such networks can be identified in some language disorders. The authors claim that these observations indicate some unexpected features of language organization that might reflect the evolution and social history of lexicons and the origins of their flexibility and combinatorial nature.

Steyversa and Tenenbaumb [9] present statistical analysis of the large-scale structure of 3 types of semantic networks: word associations, wordnet, and rogets thesaurus. This research shows that they have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering. . The authors describe a simple model for semantic growth, in which each new word or concept is connected to an existing network by differentiating the connectivity pattern of an existing node. This model generates appropriate small-world statistics and power-law connectivity distributions, and also suggests one possible fundamental basis for the effects of learning history variables on behavioral performance in semantic processing tasks.

Markosova[10] revises recent studies of syntactical word web. He presents a model of growing network in which processes such as node addition, link rewiring and new link creation are taken into account. The author argues that this model is a satisfactory minimal model explaining measured data.

Freeman and Barnett, [11] try to identify, which characteristics of the language are used in the written messages sent to the employees in a manufacturing company of medical equipment. Authors of the framework focus their research on organizational culture.

Coronges [12] uses a network analysis approach to provide information that helps to compare structural indexes of associative organization of two populations varying in age and city location; associative connections between words reveal the organization of concepts in these populations.

Brasethvik and Atle [13] present an approach to semantic document classification and retrieval based on natural language analysis and conceptual modeling. A conceptual domain model is used in combination with linguistic tools to define a controlled vocabulary for document collection.

Our research studies the structure of the company language by means of the network science and gets interesting conclusions. The investigation shows that some features are common among organizations.

III. Analysis Method

We analyze chairman's letters of annual reports for the following companies and periods: TELEFÓNICA (2009, 2008, 2007, 2006), AMPER (2009, 2008, 2007, 2006), INDRA (2009, 2008, 2007, 2006), VODAFONE (2009,2008, 2007, 2006) and JAZZTEL (2009, 2008, 2005, 2004). Chairman's letters are written in Spanish.

A. Design of word network

A text can be plotted as a graph $G = (U, L)$; where $U = \{i\}_{\ (i=1,\ldots,N)}$ is the set of N words and $L = \{i, j\}$ is the set of connections between them. Adjacent words are defined as a couple i, j, which pertains to G , and where a binary relation or link exists.

There are several assumptions made, including: repeated words correspond to the same node i, word network is not case sensitive, and the words before and after a score are not neighbours in the network.

B. Words' co-ocurrence

Table 1 shows the words with a percentage of co-occurrence bigger than 1 % for the five companies studied. It should be noted that words in English are written in italics, different words in Spanish can be translated to the same word in English, and those words in bold are repeated. Prepositions, conjunctions, articles and adjectives have the higher percentage. One can notice that some words appear in more than one company's reports, for instance *"millones"* (*"millions"*) in JAZZTEL and TELEFÓNICA; *"compañía"* (*"company"*) in AMPER and TELEFÓNICA , *"desarrollo"* (*"development"*); in INDRA and VODAFONE. The company's own name appears in letters from AMPER, INDRA, JAZZTEL and TELEFÓNICA.

A matrix B of $5x5$ dimension can be built as a representation of the coincidence of words between companies. The element B_{tq} is the coincidence percentage (%) in the texts between t and q companies. $t, q = 1$ for TELEFÓNICA; $t, q = 2$ for AMPER; $t, q = 3$ for INDRA; $t, q = 4$ for JAZZTEL; and $t, q = 5$ for VODAFONE. We show below that the main similarities are: TELEFÓNICA and AMPER in 60.71%, AMPER and INDRA in 86.36 %, and JAZZTEL and VODAFONE in 77.78 %.

$$B = \begin{pmatrix} 100 & 60.71 & 57.14 & 39.29 & 50 \\ 60.71 & 100 & 86.36 & 40.91 & 63.64 \\ 57.14 & 86.36 & 100 & 32 & 64 \\ 39.29 & 40.91 & 32 & 100 & 77.78 \\ 50 & 63.64 & 64 & 77.78 & 100 \end{pmatrix}$$

C. Structural parameters calculation

1) General characteristics: We have calculated the average distances between nodes ($< l >$), average degree ($< k >$) and the most distant vertices (d) in the word networks. These parameters are similar to those of other complex technological networks ([11], [14], [15]) and close to the value obtained in random networks, where the small world property has been likewise reported [15]. We show these characteristics in Table 2 and Table 3.

 ISBN: 978-1-61208-242-4

Table I
CO-OCCURRENCE OF WORDS IN AMPER, INDRA, JAZZTEL, TELEFÓNICA AND VODAFONE COMPANYS

	Words (%)
AMPER	**de,** ***of*** (14.07 %); **la,** ***the*** (4.55 %); **que,** ***that*** (4.22 %); **el,** ***the*** (2.93 %); **las,** ***the*** (1.79 %); **a,** ***to***(4.22 %); **ha,** ***has*** (2.05 %); **en,** ***in*** (4.09 %); **y,** ***and*** (3.58 %); **un,** ***a*** (2.44 %); **con,** ***with*** (2.81 %) ; **del,** ***of*** (2.60 %); **los,** ***the*** (2.28 %);**nos,** ***us*** (1.30 %); **por,** ***for*** (2.28 %); **para,** ***to*** (1.02 %) ; **una,** ***a*** (1.95 %); **nuestra,** ***our*** (1.14 %); **nuestro,** ***our*** (1.14 %); AMPER (1.28); **euros,** ***euros*** (1.14 %); **compañia,** ***company*** (1,46 %);
INDRA	**de,** ***of*** (11.34 %); **y,** ***and*** (8.96 %); **en,** ***in*** (7.31 %); que, *that* (5.60 %); **la,** ***the*** (4.78 %); **los,** ***the*** (3.80 %); **a** ***to*** (3.36 %); **los,** ***the*** (3.02 %); **el,** ***the*** (2.92 %); **un,** ***a*** (2.05 %); **nuestro,** ***our*** (1.79 %); INDRA (1.79 %); **con,** ***with*** (1.75 %); **las,** ***the*** (1.34 %); **nuestra,** ***our*** (1.34 %) ; futuro, *future* (1.01 %); **del,** ***of*** (1.46 %); más,*more* (1.46 %); **una,** ***a*** (1.34 %) ; **nuestros,** ***our***(1.19 %); ha, *has* (1.19 %);**desarrollo,** ***development*** **%** (1.17 %); **para,** ***to*** (1.17 %); **por,** ***for*** (1.12 %)
JAZZTEL	**de,** ***of*** (15.17 %); a, *to* (3.48 %) ; **y,** ***and*** (3.62 %); **el** (3.62 %); JAZZTEL (2.84 %); **la,** ***the*** (2.84 %); **con,** ***with*** (2.28 %); **un,** ***a*** (1.74 %); **para,** ***to*** (1.74 %); **euros,** ***euros*** (1.74 %); **se,** ***is*** (1.52 %); **por,** ***for*** (1.45 %); **una,** ***a*** (1.42 %); **clientes,** ***clients*** (1.14 %); mercado, *market* (1.45 %); **millones,** ***millions*** (1.14 %); (1.14 %); equipo, *team* (1.14 %); **nuestros,** ***our*** (1.05 %); **nos,** ***us*** (1.05 %)
TELEFÓNICA	**de,** ***of*** (22.16 %); **en,** ***in*** (9.06 %); **y,** ***and*** (7.48 %); un, *a* (3.17 %); **del,** ***of*** (2.88 %); **con,** ***with*** (2.87 %); **los,** ***the*** (2.50 %); **una,** ***a*** (2.37 %); **se,** ***is*** (2.30 %); **por,** ***for*** (2.19 %); TELEFÓNICA (2.15 %); **ha,** ***has*** (1.73 %); como, *like* (1.69 %); es, *is* (1.58 %); al, *to the* (1.58 %); **compañia,** ***company*** (1.43 %); crecimiento, *growth* (1.29 %); **millones,** ***millions*** (1.29 %); resultados, *results* (1.19 %); sector, *area* (1.18 %); no, *not* (1.01 %); **clientes,** ***clients*** (1.01 %)
VODAFONE	**de,** ***of*** (13.57 %); **y,** ***and*** (6.58 %); **en,** ***in*** (5.91 %) ; **el,** ***the*** (4.80 %); la, *the* (3.60 %); **que,** ***that*** (3.29 %);**los,** ***the*** (2.80 %); a, *to* (2.06 %); actuaciones, *actions* (2.06 %); **un,** ***a*** (1.87 %); más, *more* (1.82 %); **por,** ***for*** (1.36 %); **para,** ***to*** (1.36 %); **desarrollo,** ***development*** (1.23 %) ; **del,** ***of*** (1.20 %); **las,** ***the*** (1.20 %) ;este, *this* (1.20 %); **una,** ***a*** (1.12 %); servicios, *services* (1.12 %);

Table II
GENERAL CHARACTERISTICS IN TELEFÓNICA, AMPER, INDRA, JAZZTEL AND VODAFONE COMPANIES

		2009	2008	2007	2006	2005	2004
TELEFÓNICA	Density	0.0070947	0.0076476	0.0094565	0.0080372		
	$< k >$	4.91	4.55	4.79	4.48		
	$< l >$	3.36	3.35	3.28	3.43		
	d	10	9	8	9		
AMPER	Density	0.0112201	0.0159254	0.0115064	0.0092859		
	$< k >$	3.98	3.25	3.53	3.62		
	$< l >$	3.61	3.78	3.63	3.498		
	d	10	11	11	8		
INDRA	Density	0.0112644	0.0111099	0.0122924	0.0146433		
	$< k >$	3.83	3.73	3.69	3.90		
	$< l >$	3.52	3.64	3.55	3.51		
	d	10	11	9	8		
JAZZTEL	Density	0.0166325	0.0202877			0.0132935	0.0128408
	$< k >$	3.48	2.76			3.48	3.67
	$< l >$	3.56	4.23			3.63	3.82
	d	8	10			7	9
VODAFONE	Density	0.0145563	0.0159402	0.0139675	0.0124659		
	$< k >$	3.52	3.49	3.72	3.10		
	$< l >$	3.60	3.57	3.41	4.07		
	d	9	9	8	10		

2) Betweenness: The betweenness b_i of a node i in a network is related to the number of times that such node is a member of the set of shortest paths, which connect all the pairs of nodes in the network. If g_{nl} is the total number of possible paths from *n* to *l* nodes, and g_{nil} is the number of paths from *n* to *l* passing through *i*, then g_{nil}/g_{nl} is the proportion of paths from *n* to *l* that pass through *i*. The betweenness for a node *i* is defined as: $b_i = g_{nil}/g_{nl}$.

The functional relevance of the betweenness centrality b_i of a node is based on the observation that a node located on the shortest path between two other nodes has a greater influence over the information transfer between them. The highly connected nodes (hubs) must have high-betweenness values because there are many nodes directly and exclusively connected to these hubs and the shortest path between these nodes goes through these hubs.

Table III
GENERAL CHARACTERISTICS IN RANDOM NETWORKS

		2009	2008	2007	2006	2005	2004
TELEFÓNICA	$< k >$	4.91	4.55	4.79	4.48		
	$< l_{rand} >$	4.21	4.24	4.07	4.40		
	d_{rand}	8	10	9	9		
AMPER	$< k >$	3.98	3.25	3.53	3.62		
	$< l_{rand} >$	4.18	4.21	4.85	4.71		
	d_{rand}	9	9	12	10		
INDRA	$< k >$	3.83	3.73	3.69	3.90		
	$< l_{rand} >$	4.35	4.21	4.65	4.31		
	d_{rand}	11	9	11	10		
JAZZTEL	$< k >$	3.48	2.76			3.48	3.67
	$< l_{rand} >$	4.60	4.65			4.54	4.44
	d_{rand}	10	10			10	9
VODAFONE	$< k >$	3.52	3.49	3.72	3.10		
	$< l_{rand} >$	4.45	4.37	4.29	4.73		
	d_{rand}	9	9	8	11		

In all studied networks, values for betweenness b are ranged over several orders of magnitude and there are low-connectivity and high-connectivity nodes, which exhibited a wide range of betweenness values. It is shown in Figure 1 for TELEFONICA and AMPER company in 2009, where betweenness, b, is plotted as a function of connectivity (degree), k. This result indicates the existence of a large number of nodes with high betweenness but low connectivity. Although the low connectivity of these words would imply that they are unimportant, their high betweenness suggests that these words may have a global impact. From a topological point of view, these words are positioned to connect regions of high clustering (containing hubs), even though they have low local connectivity. The existence of such words points to the presence of modularity in the network, and therefore suggests that these words may represent important connectors that link these modules. This behaviour was also found in the yeast protein interaction network [16].

Articles, pronouns, prepositions, conjunctions, adverbs and adjectives appear in the highest positions in the ranking and these words are similar in all texts. The first appearance of verbs and sustantives is lower in the ranking for all reports.

Depending on the year, in TELEFÓNICA and AMPER companies, the first ranked verb is *"ser"* (*"to be"*) or *"haber"* (*"to have"*), in JAZZTEL company the verb *"dudar"*, (*"to hesitate"*) also appears in that position; the first verb is *"haber"* (*"to have"*) in INDRA company and in VODAFONE company is *"suponer"* (*"to suppose"*), *"contribuir"* (*"to contribute"*), *"haber"* (*"to have"*) and *"descargar"* (*"to download"*). The first substantive in TELEFÓNICA company is *TELEFÓNICA*; in AMPER company the first nouns are *"mercado"* (*"market"*), *AMPER* and *"año"* (*"year"*); whereas in JAZZTEL company is

Table IV
DECREASING RANKING REGARDING TO BETWEENNESS IN TELEFÓNICA'S REPORT

Rank	2009	2008	2007	2006
1	de *of*	de *of*	de *of*	de *of*
2	la *the*	en *in*	y *and*	en *in*
3	que *that*	y *and*	el *the*	y *and*
4	en *in*	la *the*	en *in*	el *the*
5	y *and*	que *that*	la *the*	la *the*
6	el *the*	un *a*	del *of*	que *that*
7	a *to*	los *the*	que *that*	un *a*
8	los *the*	el *the*	un *a*	a *to*
9	las *the*	TELEFÒNICA	a *to*	las *the*
10	con *with*	del *of*	para *to*	por *for*
11	una *a*	a *to*	los *the*	una *a*
12	un *a*	con *with*	las *the*	nuestra *our*
13	para *to*	las *the*	con *with*	los *the*
14	TELEFÓNICA	para *to*	TELEFÓNICA	TELEFÓNICA
15	del *of*	por *for*	clientes *clients*	del *of*
16	su *its*	más *more*	crecimiento *growth*	más *more*
17	se *is*	su *its*	una *a*	con *with*
18	ha *has*	es *is*	al *to the*	es *is*
19	es *is*	...	2007	se *is*
20	más *more*	...	por *for*	crecimiento *growth*
21	al *to the*	...	como *like*	para *to*
...	...	...	...	...
27	...	...	ha *has*	
...	...	...	...	...
35	...	...	...	ha *has*
...	...	...	...	...

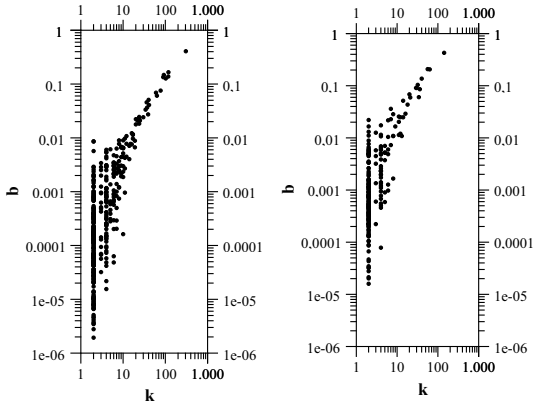


Figure 1. Degree (k) versus betweenness (b) plotted in logarithmic scale for 2009 reports in TELEFÓNICA company (left) and AMPER company (right)

"premios" (*"awards"*), *JAZZTEL* and *"mercado"* (*"market"*); in INDRA company the first sustantive is *"futuro"* (*"future"*), *"fruto"* (*"to bear fruit"*), *"confianza"* (*"confidence"*), *"soluziona"* (*without translation, project name*); and in VODAFONE company the first noun is *"actuaciones"* (*"actions"*), *"desarrollo"* (*"development"*)and *"servicios"* (*"services"*). We show these results in Table 4. It should be emphasized that words in English are written in italics and different words in Spanish can be translated to the same word in English.

3) Communities: Different methods have been developed in order to find communities in networks. Basically, these methods can be grouped as spectral methods (e.g., [18]), divisive meth-ods (e.g., [19]), agglomerative methods (e.g., [20]), and local methods (e.g., [21]). The choice of the best method depends of the specific application, including the network size and number of connections.

We have carried out a study of the community structure in the word networks by measuring the similarities between nodes by means of Walktrap Algorithm [7] [22]. Walktrap algorithm is an agglomerative approach to detect communities, which starts with a community for each node such that the number of partitions $|\rho| = n$ and build communities by amalgamation.

Walktrap method uses random walks on G to identify communities. At each step in the random walk, the walker is at a node and moves to another node chosen at random yet uniformly from its neighbors. The sequence of visited nodes is a Markov chain where the states are the nodes of G. An N x N dimension adjacency matrix $A(G)$ can be built as a bidimensional representation of the relationships between words, where $A_{ij} = 1$ when a connection between the nodes i and j exists and $A_{ij} = 0$ otherwise. At each step the transition probability from node i to node j is $P_{ij} = \frac{Aij}{ki}$, which is an element of the transition matrix P for the random walk. We can also write $D^{-1}A$, where D is the diagonal matrix of the degrees ($\forall_i, D_{ii} = k_i$ and $D_{ij} = 0$ where $i \neq j$. The random walk process is driven by powers of P: the probability of going from *i* to *j* in a random walk of length t is $(P^t)_{ij}$, which we will denote simply as P^t_{ij} . All of the transition probabilities related to node *i* are contained in the i^{th} row of P_t denoted as $P^t_{i\bullet}$. Then we define an inter-node distance measure as:

$$s_{ij} = \sqrt{\sum_{q=1}^{n} \frac{(P^t_{iq} - P^t_{jq})^2}{k_q}} = \| D^{1/2}P^t_{i\bullet} - D^{1/2}P^t_{j\bullet} \| \quad (1)$$

where $\| \bullet \|$ is the Euclidean norm of R^n . This distance can also be generalized as a distance between communities: s_{CiCj} or as a distance between a community and a node: s_{Cij}

We then use this distance measure in our algorithm. The algorithm uses an agglomerative approach, beginning with one partition for each node ($|\rho| = n$) . We first compute the distances for all adjacent communities (or nodes in the first step). In each step α , two communities are chosen based on the minimization of the mean σ_α of the squared distances between each node and its community.

$$\sigma_\alpha = \frac{1}{n}\Sigma_{C_i \in \rho_\alpha}\Sigma_{i \in C_i} s^2_{iC_i} \quad (2)$$

Instead of directly calculating this quantity, we first calculate the variations $\Delta\sigma_\alpha$. Due to the fact that the algorithm uses a Euclidean distance, we can efficiently calculate these variations as

$$\Delta\sigma(C_1, C_2) = \frac{1}{n}\frac{|C_1||C_2|}{|C_1| + |C_2|} s^2_{C_1C_2} \quad (3)$$

The community merges when the lowest $\triangle\sigma$ is performed. The transition probability matrix is then updated accordingly.

$$P^t_{(C_1 \cup C_2)\bullet} = \frac{|C_1|P^t_{C_1\bullet} + |C_2|P^t_{C_2\bullet}}{|C_1| + |C_2|} \quad (4)$$

and the process is repeated again, updating the values of s and $\Delta\sigma$ and then performing the next merge. After $n-1$ steps, we get one partition that includes all the nodes of the network $\rho_n = \{N\}$. The algorithm creates a sequence of partitions $(\rho_\alpha)_{1\leq\alpha\leq n}$. Finally, we use modularity to select the best partition of the network, calculating Q_{ρ_α} for each partition and selecting the partition that maximizes modularity.

We define modularity Q as the fraction of links within communities minus the expected value of the same quantity for a random network. Let A_{ij} be an element of the networks adjacency matrix and suppose the nodes are divided into communities such that node i belongs to community C^i. Then Q can be calculated as follows:

$$Q = \frac{1}{2m}\sigma_{ij}\{A_{ij} - \frac{k_i k_j}{2m}\}\delta_{C_i C_j} \quad (5)$$

where the $\delta_{C_i C_j}$ function is 1 if $C_i = C_j$ and 0 otherwise. m is the number of links in the graph, and k_i is the degree of a node i. The sum of the term $\frac{k_i k_j}{2m}$ over all pair nodes in a community represents the expected fraction of links within that community in an equivalent random network where node degree values are preserved.

All word networks have several main communities, which have high connected nodes but with few connections to the rest of the network. These so clearly defined communities suggest a network structure. The community rank by percentage of words is shown in Table 6. In Figure 2, we plot the community rank for TELEFÓNICA's reports in 2006 and 2007. We detected 88 communities in 2006 and 42 communities in 2007 within this company. In Figure 3, we display the community rank for INDRA's reports in 2008 and 2009; 33 communities in 2008 and 39 communities in 2009 were found.

In each community, we also detected that the higher probability of appearance according to the type of word occurs among nouns, verbs and adjectives. In Figure 4, we show the percentage by the type of word in rank 1 community and in rank 2 community for INDRA's report in 2007.

4) Motifs: Network motifs are connectivity-patterns (subgraphs) that occur frequently in the networks. Most networks studied in biology, ecology, communication and others fields have been found to show a small set of network motifs; in most cases these motifs are repeated. In [22] a network motif is defined as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks"

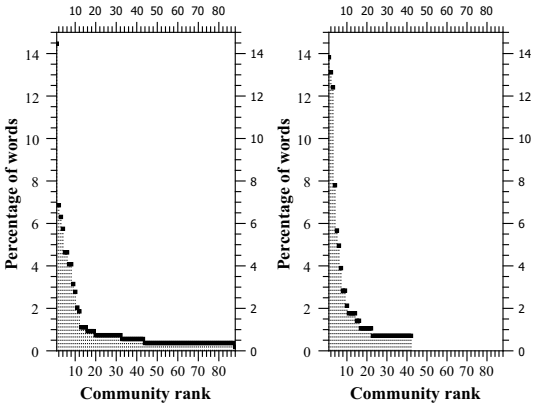


Figure 2. Left, community rank by percentage of words over 88 communities detected for TELEFÓNICA's report in 2006. Right, community rank by percentage of words over 42 communities found for TELEFÓNICA's report in 2007.

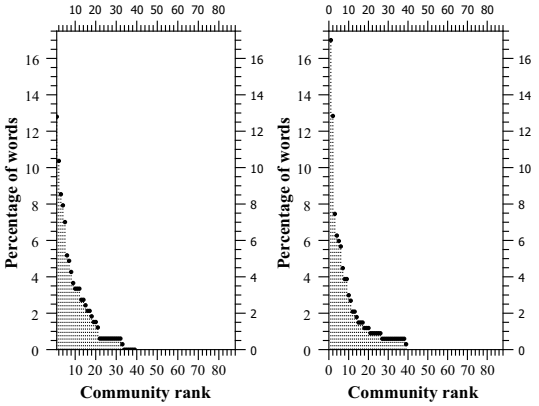


Figure 3. Left, community rank by percentage of words over 33 communities detected for INDRA's report in 2008. Right, community rank by percentage of words over 39 communities found for INDRA's report in 2009.

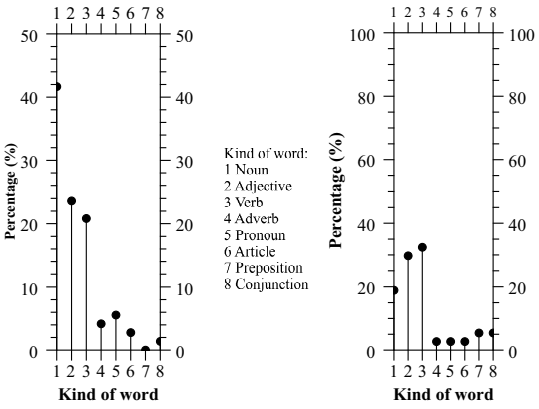


Figure 4. Percentage by kind of word in rank 1 community (left) and in rank 2 community (right) for INDRA's report in 2007.

Table V
COMMUNITY RANK BY PERCENTAGE OF WORDS IN TELEFÓNICA'S, AMPER'S, INDRA'S, JAZZTEL'S AND VODAFONE'S REPORTS

		2009	2008	2007	2006	2005	2004
TELEFÓNICA	Number of communities	83	79	42	88		
	Rank						
	1	11.35 %	10.25 %	13.83 %	14.47 %		
	2	8.59 %	9.74 %	13.12 %	6.86 %		
	3	7.71 %	8.89 %	12.41 %	6.31 %		
	4	6.55 %	6.67 %	7.80 %	...		
AMPER	Number of communities	40	21	31	22		
	Rank						
	1	18.29	16.49	13.49	16.62		
	2	14.29	15.46	8.22	15.83		
	3	11.14	13.92	7.57	12.66		
	4	9.14	7.22	6.25	11.35		
	5	6.57	6.70	...	8.44		
	6	...	...	...	6.33		
INDRA	Number of communities	39	33	30	31		
	Rank						
	1	17.01	12.80	26.03	17.62		
	2	12.84	10.37	12.33	16.86		
	3	7.46	8.54	10.96	8.43		
	4	6.27	7.93	...	6.90		
	5	...	7.01	...	...		
JAZZTEL	Number of communities	28	28			28	26
	Rank						
	1	14.57	8.40			15.32	20.14
	2	13.07	7.63			9.68	11.51
	3	12.56	6.87			7.26	9.35
	4	10.05	...			...	8.99
	5	8.54	...			...	7.91 %
	6	...	...			...	6.47
	7	...	...			...	6.12
VODAFONE	Number of communities	21	24	32	24		
	Rank						
	1	14.96	22.17	15.02	19.83		
	2	6.84	11.79	13.04	10.33		
	3	6.41	11.32	12.25	9.92		
	4	...	10.85	11.07	8.26		
	5	...	7.08	...	7.02		
	6	...	...	...	6.61		

The procedure used to detect network motifs in all word networks is described in [22]. This method samples the sub-graphs of the size n and estimates the appearances of them in the whole graph based on the frecuencies obtained in the samples. A sub-graph is sampled using a simple iterative procedure by selecting connected links until a set of n nodes is reached. The process is as follows:

L_S is the set of picked links.
N_S is the set of all nodes that are touched by the links in L_S.
L_S and N_S are inited to be empty sets.

1) Pick a random edge $l_1 = (h, i)$. Update $L_S = \{l_1\}$, $N_S = \{h, i\}$.
2) Make a list L of all neighboring links of L_S. Omit from L all links between two members of N_S. If L is empty return to 1.
3) Pick with a probability P edge $l = \{j, k\}$ from L.
4) Update $L_S = L_S \cup \{l\}, N_S = N_S \cup \{j, k\}$.
5) Repeat steps $2 - 3$ until completing n-node sub-graph S.
6) Calculate the probability P to sample S.

A sub-graph of size n-node can be represented as an adjacency matrix of n x n dimension M, where $M_{ij} = 1$ when a connection between nodes i and j exists and $M_{ij} = 0$ otherwise. For simplicity, in this study, we have symbolized this adjacency matrix as a long binary integer extracted by concatenation of its rows. This number is named Identity (Id).

We applied the method described above in all corporations and their different reports to look up 3-node and 4-node connected sub-graphs ($n = 3$ and $n = 4$). We detected two kinds of 3-node sub-graphs and six types of 4-node sub-graphs. These graphs are $Id = 78$, $Id = 238$, $Id = 4,382$, $Id = 4,698$, $Id = 4,958$, $Id = 13,260$, $Id = 13,278$ and $Id = 31,710$; in all company reports. In Table 6, we show the relationship between Id and M for them.

Table VI
RELATIONSHIP BETWEEN Id AND M FOR 3-NODE AND 4-NODE SUB-GRAPHS, WHICH HAVE BEEN DETECTED IN THE ANNUAL REPORTS FOR ALL COMPANIES

	Id	M		Id	M
3 node-sub-graph	78	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	4-node sub-graph	4,382	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
	238	$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$		4,698	$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
				4,958	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
				13,260	$\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$
				13,278	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$
				31,710	$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$

In the Figures 5 and 6, we show sub-graphs that have been found and their appearances for all analyzed companies in 2009. For instance, in TELEFÓNICA company the following appearances of 3-node graphs were detected: $40,393$ with $Id = 78$; 308 with $Id = 238$ and the following identifiers for 4-node graphs were found: $1,541,382$ with $Id = 4,382$; $255,232$ with $Id = 4,698$; $45,702$ with $Id = 4,958$; $2,079$ with $Id = 13,260$; $1,404$ with $Id = 13,278$; 23 with $Id = 31,710$

Uniqueness in a graph is the number of times it appears with completely disjoint groups of nodes. For all companies in their different reports, we have also calculated the graph's uniqueness percentage over its total appearances and this parameter is higher in the graphs with $Id = 238$, for AMPER, JAZZTEL and VODAFONE companies, and in the graph with $Id = 31,710$ for TELEFÓNICA and INDRA companies. In Figures 7 and 8, we show the percentage uniqueness in each sub-graph of the reports for all corporations in 2009. So, for $Id = 238$ in AMPER this uniqueness percentage is 29,17% , in JAZZTEL it's 34,78% and in VODAFONE it's 12,20%; for $Id = 31,710$ in TELEFÓNICA it's 8,70% and in INDRA it's 50,00%. We observe that the percentage uniqueness is greater in the sub-graphs with smaller appearances.

These results suggest common communication characteristics in business language.

We also estimate that the average appearances in random networks are 1,000 nodes. The method used to generate

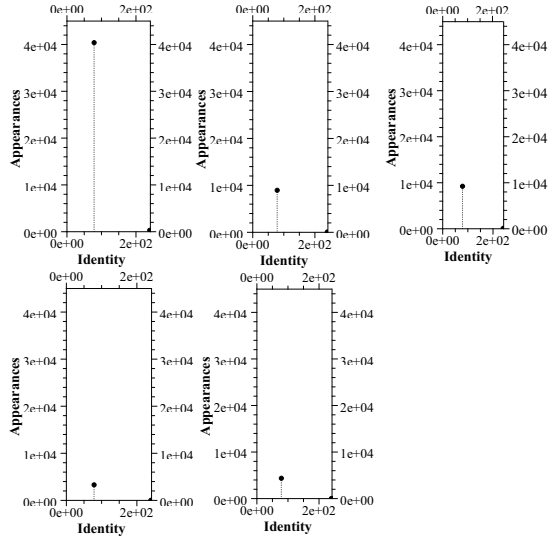


Figure 5. Appearances by Id for 3-node sub-graphs in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

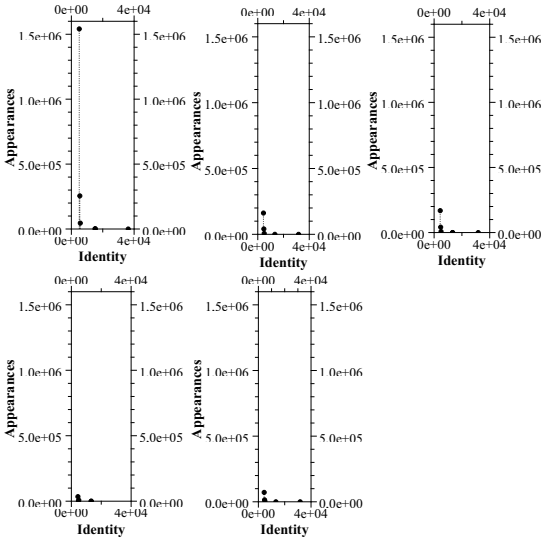


Figure 6. Appearances by Id for 4-node sub-graphs in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

random networks is the switching mechanism where we switch between links while keeping the number of incoming links of each node of the real network. The number of switches is a random number within the range of 100-200 times that the number of links appear in the network. In random networks the obtained results are also in good agreement with the real networks. If we consider the number of average appearances, the most frequent 3-node sub-graph is also: $Id = 78$ and the 4-node sub-graph with most appearances is also: $Id = 4,382$ and $Id = 4,698$, as it is shown in Figures 9 and 10 for reports in 2009.

IV. Conclusions and Future Work

In this paper, through the use of a large amount of data, we have been able to examine the syntactical language structure used by several information technology companies theoretically. We have checked their annual reports and measured various properties: average degree, main shortest path and betweenness. Additionally, we detected communities and motifs. There are common properties but other characteristics appear in some corporations only.

All company reports show a small world property which is a characteristic of complex systems. The parameters average degree ($< k >$) and the most distant vertices (d) are also similar.

Furthermore, TELEFONICA and AMPER, AMPER and INDRA show high coincidence in words. JAZZTEL and VODAFONE also have many common words. This result can suggest a common essence in the companies.

All word networks also have main communities which shows a high hierarchy among each network. In each community, the probability of appearance according to the type of word is higher for nouns, verbs and adjectives.

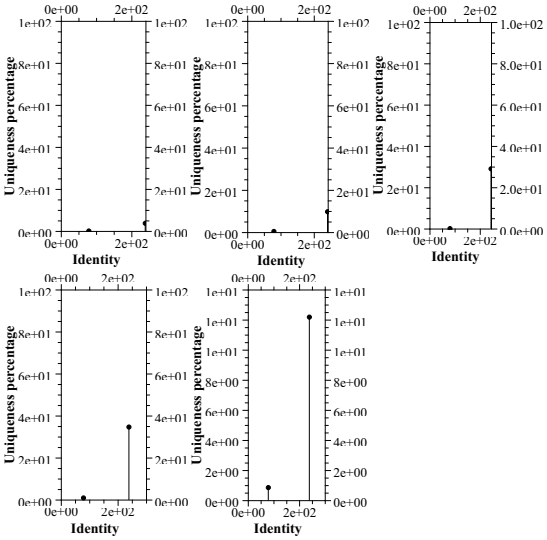


Figure 7. Uniqueness percentage for 3-node sub-graphs in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

In all corporation reports there are a large number of nodes with high betweenness but low connectivity; although the low connectivity of these words would imply that they are unimportant, their high betweenness suggests that these words have a global impact.

From a structural point of view, all companies have 3-node and 4-node common connected sub-graphs. The more frequent graph types are those with Id. 78, 4382, 4698 and 4958.

 ISBN: 978-1-61208-242-4

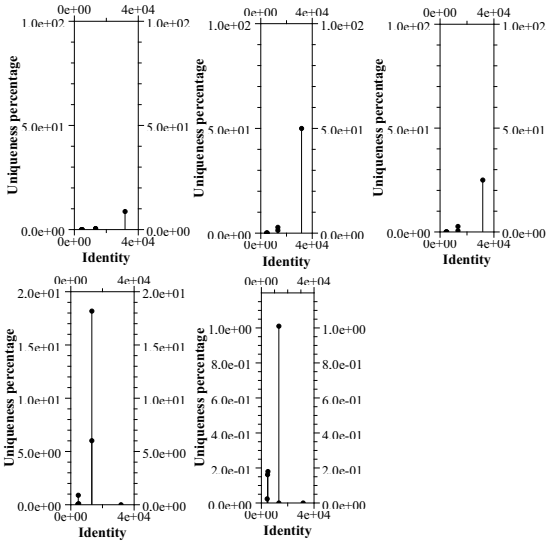


Figure 8. Uniqueness percentage for 4-node sub-graph in 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-center).

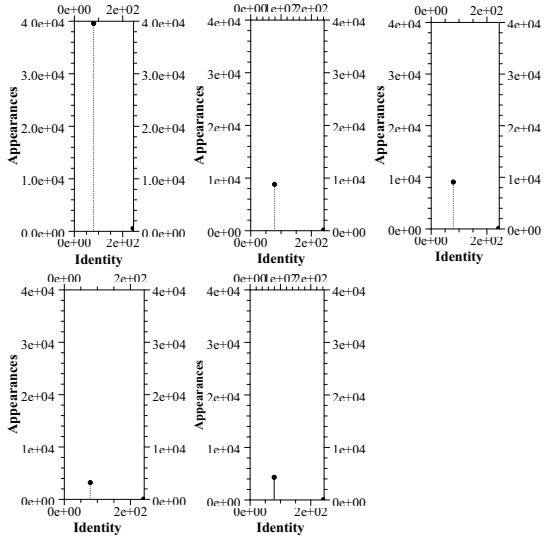


Figure 9. Appearances for 3-node sub-graphs in random networks for 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-right).

This study about the syntactic structure of the language in the organization reports, can help to identify specific and common properties in the companies. In future works, we improve this research by means of a semantic analysis.

REFERENCES

[1] E.R. Kandel, Principles of neural science, McGraw-Hill, Health Professions Division, 2000.

[2] M. Faloutsos , P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology", SIGCOMM: Conference on Communication IEEE,. 1999, pp. 251-262.

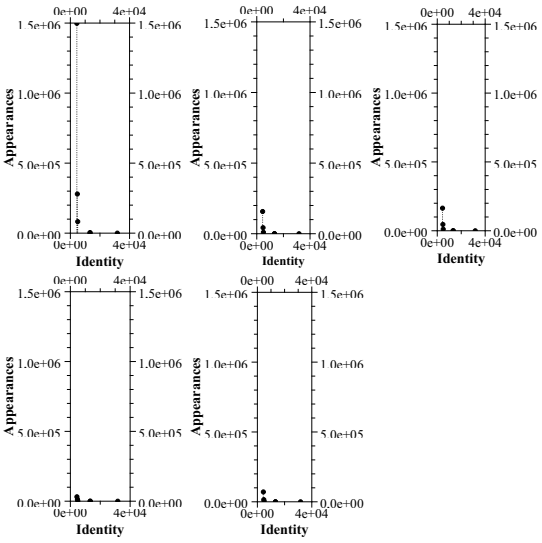


Figure 10. Appearances 4-node sub-graphs in random networks for 2009 reports, TELEFÓNICA (upper-left), INDRA (upper-center), AMPER (upper-right), JAZZTEL (bellow-left) and VODAFONE (bellow-right).

[3] J. Spencer, D. Johnson , A. Hastie, and L. Sacks, "Emergent properties of the BT SDH network", BT Technology Journal, vol. 21, 2003, pp. 28-36.

[4] M.L. Mouronte et al., "Complexity in Spanish optical fiber and SDH transport networks", Computer Physics Communications, vol. 180, 2009, pp. 523-526.

[5] K. Brner, S. Sanyal, and A. Vespignani, Network science, Annual Review of Information Science & Technology, vol. 41, B. Cronin, ed., pp. 537607.

[6] R. Milo, et al., Network Motifs: Simple Building Blocks of Complex Networks, Science vol. 298, 2002, pp. 824-827.

[7] P. Pons, and M. Latapy, "Computing communities in large networks using random walks", ISCIS2005, 2005, pp. 284-293.

[8] R. Ferrer i Cancho, R. and R. V. Sole, "The small world of human language", Proc. R. Soc. Lond. B., vol. 268, 2001, pp. 2261-2265.

[9] M. Steyversa and J.B. Tenenbaumb, "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth", Cognitive Science, vol. 29, No. 4178, 2005, pp. 4178.

[10] M. Markosova, "Network model of human language," Physica A , vol. 387, 2008, pp. 661-666.

[11] C.A. Freeman and G.A. Barnett, "An alternative approach to using interpretative theory to examine corporate messages and organizational culture", L. Thayer and Barnett G.A. (ed.): Organization Communication. Emerging Perspectives, Norwoord, New Jersey: Ablex, vol. 4, 1994.

[12] K.A. Coronges, "Structural Comparison of Cognitive Associative Networks in Two Populations",Journal of Applied Social Psychology, vol. 37, No. 9, 2005, pp. 2097–2129.

[13] T. Brasethvik and J. Atle, ”Natural Language Analysis for Semantic Document Modeling,” http://www.idi.ntnu.no/ brase/pub/nldb-brase-gullaV40.pdf, 2011.

[14] M.E.J. Newman, ”The Structure anf Functions of Complex Networks”, SIAM Review, vol 45, No. 2, 2003, pp. 167-256.

[15] S.N. Dorogovtsev and J.F.F. Mendes, ”Evolution of Networks. From Biological Nets to the Internet and WWW”, Oxford University Press, 2003.

[16] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, ”High-Betweenness Proteins in the Yeast Protein Interaction Network”, Journal of Biomedicine and Biotechnology, vol. 2, 2005, pp. 96-103.

[17] S. Shen-Orr1, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, ”Network Motifs: Simple Building Blocks of Complex Networks”, vol. 298, No. 2002, pp. 824-827.

[18] M. Newman, ”Modularity and community structure in networks”, Proceedings of the National Academy of Sciences of the United States of America , vol. 103, No. 23, 2006, pp. 8577-8582.

[19] M. Girvan and M. Newman, ”Community structure in social and biological networks”, Proceedings of the National Academy of Sciences of the United States of America, vol. 99, No. 12, 2002, pp. 7821-7826.

[20] A. Clauset, M. E. J. Newman, and C. Moore, ”Finding community structure in very large networks”, Physical Review E, 70, No. 6, 2004.

[21] A. Clauset, ”Finding local community structure in networks”, Physical Review E, Vol. 72, No. 4, 2005.

[22] B. Fields, et al., Analysis and Exploitation of Musician Social Networks for Recommendation and Discovery, IEEE Transactions on Multimedia, vol. 13, 2011, pp. 674-686.

[23] ”AMPER”, http://www.amper.es/index.cfm?lang=sp, July 2012.

[24] ”JAZZTEL”, http://www.jazztel.com/home, July 2012.

[25] ”VODAFONE”, http://www.vodafone.es/particulares/es/?cid=12072012-000000148, July 2012.

[26] ”INDRA”, http://www.indracompany.com/, July 2012.

[27] ”TELEFÓNICA”, http://www.telefonica.com/es/home/jsp/home.jsp, July 2012.

 ISBN: 978-1-61208-242-4

How to Find Important Users in a Web Community? Mining Similarity Graphs

Clemens Schefels
Institute of Computer Science, Goethe-University Frankfurt am Main
Robert-Mayer-Straße 10, 60325 Frankfurt am Main
Email: schefels@dbis.cs.uni-frankfurt.de

***Abstract*—In this paper we provide a useful tool to the web site owner for enhancing her/his marketing strategies and rise as consequence the click rates on her/his web site. Our approach addresses the following research questions: which users are important for the web community? Which users have similar interests? How similar are the interests of the users of the web community? How is this specific community structured? We present a framework for building and analyzing weighted similarity graphs, e.g., for a social web community. For that, we provide measurements for user equality and user similarity. Furthermore, we introduce different graph types for analyzing profiles of web community users. We present two new algorithms for finding important users of a community.**

***Keywords*-Computer aided analysis; World Wide Web; Data analysis; Graph theory;**

I. INTRODUCTION

Nowadays, web-based user communities enjoy great popularity. Facebook[1] has more than 900 million members and even the relatively new Google+[2] about 170 million. In this highly competitive environment, it is crucial for web site owners to understand and satisfy their web community.

Previous research discovered community structures in these networks, but focused only on the pure friendship structure of these communities [1]. In this paper, we present a tool for building and mining similarity graphs. These similarity graphs are built from the interest profiles of the users of a web community. We use the Gugubarra framework [2], [3], developed by DBIS at the Goethe-University Frankfurt, to build interest profiles of web users.

In Gugubarra each user profile is stored as a vector that presents the supposed interests of a user u_m related to a topic T_i at time t_n. Each vector row contains the calculated interest value of the user for a given topic. The values of the interest are between 0 and 1, while 1 indicates high interest and 0 indicates no interest for a topic (see Figure 1). Gugubarra generates for each registered user several profiles:

A *Feedback Profile*, (FP), which stores the data explicitly given by a user. For that, we ask the users from time to time about their interests in respect to a set of predefined topics.

[1]https://www.facebook.com/
[2]https://plus.google.com/

A *Non-Obvious Profile*, (NOP), which stores behavioral data not explicitly given by the user, but automatically created by analyzing the user behavior on the web site. The behavioral data stored in the NOP indicates, for example, which pages a user has visited, and which actions she/he has performed on that web page. Most of this information is extracted out of the web server log, but Gugubarra has refined the common click-stream analysis [4], [5], by extending it with new concepts, namely: *zones*, *topics*, *actions*, and *weights* [3], [6].

In [7], we introduced the *Relevance Profile* (RP). An RP is calculated by integrating the two available profiles for the user, the NOP and the FP. The benefit of the RP is that it integrates both, calculated data as well as explicit feedback of the user, in a flexible way into one single user profile. Figure 1 shows an example of an RP, where we calculated the data of a user u_m at time t_n, based on her/his behavior and explicit feedback, showing a supposed low interest in topic T_1 (0.3), high interest in topic T_2 (1.0), and no interest in topic T_3 (0.0).

$$RP_{u_m,t_n} = \begin{pmatrix} 0.3 \\ 1.0 \\ 0.0 \end{pmatrix} \begin{matrix} \leftarrow T_1 \\ \leftarrow T_2 \\ \leftarrow T_3 \end{matrix}$$

Figure 1. Relevance Profile of user u_m for three topic T_1, T_2, T_3.

In what follows, we assume that users are aware and have granted permission that implicit data is collected and kept in their profile for them.

To measure the similarity of the users we are using different techniques from graph theory. First, we will introduce the similarity threshold that helps the web site owner in building the graphs of her/his community. Second, we will provide several algorithms to find important users in the similarity graph. There exists not only one valid definition for importance of users because it depends–as always–on the point of view. For this reason, we provide nine algorithms to discover the importance of users. Two of these algorithms are new designed in respect to the needs of similarity graphs.

The rest of the paper is structured as follows: Section II recalls the basic concepts that will be used in the rest of

 ISBN: 978-1-61208-242-4

the paper. In Section III, we define the similarity of users. Section IV presents the main contribution of this paper, our analysis tool for building and mining similarity graphs. In Section V we use our analysis tool with a real usage dataset. Section VI presents the conclusions of this work.

II. Literature Review

In this section we introduce basic concepts from literature that are used in our framework.

A. Similarity measurement

Due to the fact that the RP contains all information about the interests of the users, we want to use it to compute the similarity between the interests of *all* users. First we have to definite the equality of users:

Two users u_i and u_j are equal in respect to a topic T_r of a web site at time t_n if the interest values of T_r of their RPs are equal:

$$RP_{u_i,t_n}(T_r) = RP_{u_j,t_n}(T_r) \; where \; i \neq j. \tag{1}$$

To compare users we need a measurement for similarity. Similarity measurements are very common in the research field of data mining. For example, documents are often represented as feature vectors [8], which contain the most significant characteristics like the frequency of important keywords or topics. To compute the similarity of documents, the feature vectors are compared with the help of distance measurements: the smaller the distance the more similar the documents are.

Gugubarra interest profiles, i.e., the RP, can be considered as feature vectors of the users, too. They contain the most significant characteristics of our users, e.g., the interests in different topics of a web site. Therefore we can use the similarity measurements of data mining theory to compute similarity between the members of our community.

An important requirement on the similarity measurement algorithm is its performance, because a web community can cover lots of users. Consequently we have to choose a similarity measurement with a high performance so that the analysis program will scale with the high number of users. Aggarwal et al. proved in [9] that the *Manhattan Distance*, also known as *City Block Distance* or *Taxicab Geometry*, is very well suited for high dimensional data. We shared in [6] that web sites may have up to 100 topics. Thus, we have to deal with very high dimensional feature vectors, i.e., one dimension per topic.

The Manhattan Distance (L_1-norm) [10] is defined as follows:

$$d_{\text{Manhattan}}(a,b) = \sum_i |a_i - b_i| \tag{2}$$

with $a = RP_{u_m,t_n}$, $b = RP_{u_r,t_n}$ and $m \neq r$.

B. Graph Theory

Leonhard Euler founded the graph theory with the Seven Bridges of Königsberg problem [11]. In this section, we present the basic definitions of graph theory that are necessary for our tasks.

A *graph* G [12] is a tuple $(V(G), E(G))$. $V(G)$ is a set of *vertices* of the graph and $E(G)$ is the set of *edges* which connects the vertices[3].

A graph G can be represented [14] by an *adjacency matrix* $A = A(G) = (a_{ij})$. This $n \times n$ matrix, n is the sum of the vertices of G, is defined as follows:

$$a_{ij} = \begin{cases} 1 & \text{if } \{v,w\} \in E(G) \\ 0 & \text{otherwise.} \end{cases} \quad with \; v, \; w \in V(G) \tag{3}$$

In a *simple graph* an edge connects always *two* vertices [15]. This means that $E(G)$ consists of unordered pairs $\{v,w\}$ with $v,w \in V(G)$ and $v \neq w$ [12]. In a social network vertices could represent the members of this network and the edges could stand for the friendship relation between these vertices–so friends are connected together.

Every pair of distinct vertices of a *complete graph* [12] are connected together.

The connections between edges can be *directed* or *undirected*. In a directed graph the edges are an ordered pair of vertices v, w and can only be traversed in the direction of its connection. This means that a *simple graph* is undirected. This feature is very useful, e.g., to model the news feed subscriptions of a user in a social network, a one-way friendship.

A *loop* is a connection from a vertex to itself [14]. A loop is not an edge.

Labeled vertices make graphs more comprehensible. Vertices can be labeled with identifiers, e.g., in the social network graph with the names of the users.

In the same way edges can be labeled to denote the kind of connection. In the social network graph example, the label could represent the kind of relation between users, e.g., friend or relative.

With *weighted graphs*, the strength of the connection between the single vertices can be modeled. Every edge has an assigned weight. In a social network the weight could be used to display the degree or importance of the relationship of the users. A weighted graph can also be represented by an adjacency matrix (see Definition 3 above) where a_{ij} is the weight of the connection of $\{v,w\}$. See Example 4 for an adjacency matrix of a similarity graph of five users:

[3]Sometimes it is postulated [12] that $V(G)$ and $E(G)$ has to be finite but there exists also definitions about infinite graphs [13]. However, the number of web site users should be finite.

 ISBN: 978-1-61208-242-4

$$A = \begin{pmatrix} 0.00 & 1.28 & 1.19 & 2.79 & 1.18 \\ 1.28 & 0.00 & 1.63 & 2.83 & 1.90 \\ 1.19 & 1.63 & 0.00 & 2.50 & 1.35 \\ 2.79 & 2.83 & 2.50 & 0.00 & 2.85 \\ 1.18 & 1.90 & 1.35 & 2.85 & 0.00 \end{pmatrix} \quad (4)$$

Every number represents the weight of the edges between two vertices, e.g., $a_{2,4} = 2.83$ represents the edge weight of the two vertices with the numbers 2 and 4. The diagonal of this matrix is 0.00 because the graph has no loops. In an undirected graph the adjacency matrix is symmetric.

A vertex w is a *neighbor* of vertex v if both are connected via the same edge. The neighborhood of v consists of *all* neighbors of v. In a social network a direct friend is a neighbor and all direct friends are the neighborhood.

A *path* [16] through a graph G is a sequence of edges $\in E(G)$ from a starting vertex $v \in V(G)$ to an end vertex $w \in V(G)$. If there exists a path form vertex v to w both vertices are connected. The number of edges on this path is called *length* of the path and the *distance* between v and w is the length of the shortest path between these two vertices. A path with the same start and end point is called *cycle*. Two vertices v and w are *reachable* from each other if there exists a path with the start point v and the end point w. If all vertices are reachable from every vertex the graph is called *connected*.

G' is a *subgraph* [14] of G if $V(G') \subset V(G)$ and $E(G') \subset E(G)$. G is than the *supergraph* of G' with $G' \subset G$.

A *community* in a graph is a *cluster* of vertices. The vertices of a community are dense connected.

C. Importance

There exist many algorithms to measure the importance of a vertex in graph. We introduce seven of the most common algorithms:

Sergin Brin and Lawrence Page [17] used their *PageRank* algorithm to rank web pages with the link graph of their search engine Google[4] by importance. This algorithm is scalable on big data sets (i.e., search engine indices). Usually the PageRank algorithm is for unweighted graphs. But there exists also implementations for weighted graphs [18]. Pujol et al. [19] developed an algorithm to calculate the reputation of users in a social network. The results of the comparison of their algorithm with the PageRank show that the PageRank is also well suited for reputation calculation, i.e., importance calculation.

The *Jaccard similarity coefficient* [20] of two vertices is the number of common neighbors divided by the number of vertices that are neighbors of at least one of the two vertices being considered [21]. Here the pairwise similarity of all vertices is calculated.

The *Dice similarity coefficient* [21] of two vertices is twice the number of common neighbors divided by the sum of the degrees of the vertices. Here the pairwise similarity of all vertices is calculated.

Nearest neighbors degree calculates the nearest neighbor degree for all vertices. In [22] Barrat et al. define a nearest neighbor degree algorithm for weighted graphs.

Closeness centrality [23] measures how many steps are required to access every other vertex from a given vertex.

Hub score [24] is defined [21] as the eigenvector of AA^T where A is the adjacencies matrix and A^T the transposed adjacencies matrix of the graph.

Eigenvector centrality [25] [21] correspond to the values of the first eigenvector of the adjacency matrix. Vertices with high eigenvector centralities are those which are connected to many other vertices which are, in turn, connected to many others.

III. User Similarity

In Gugubarra, the RP provides the most significant information about a user which is calculated from all implicit and explicit feedback profiles. To calculate user similarity we take the RP interest value of every topic of each user and calculate the Manhattan Distance between all users of the web community as illustrated in the following example:

Lets assume we have a web site with three topics T_1, T_2, and T_3. This web site has two registered users u_1 and u_2. The RPs of the two users were calculated at time t_1:

$$RP_{u_1,t_1} = \begin{pmatrix} 1.0 \\ 0.5 \\ 0.0 \end{pmatrix}, \; RP_{u_2,t_1} = \begin{pmatrix} 0.6 \\ 0.8 \\ 0.2 \end{pmatrix} \quad (5)$$

The Manhattan Distance is calculated as follows:

$$d_{\text{Manhattan}}(RP_{u_1,t_1}, RP_{u_2,t_1}) = \quad (6)$$

$$= |1.0 - 0.6| + |0.5 - 0.8| + |0.0 - 0.2| = 0.9$$

where 0.9 is the distance of the interests of the both users, i.e., the similarity.

Therefore, our focus is on a large group of users (i.e., the whole web community) and not only on a single user or on a single topic. The following sections should clarify research questions such as:

- Which users are important for the web community?
- Which users have similar interests?
- How similar are the interests of the users of the community?
- How is this specific community structured?

[4] https://www.google.com/

By answering these questions we want to give the web site owner a useful tool to enhance her/his marketing strategies and rise as consequence the click rates of her/his portal.

IV. Analysis of Similarity Graphs

We developed a new tool for building and analyzing similarity graphs. We integrated several algorithms from different research areas for the analysis of the graphs. This tool is written in R[5]. R is an open source project with a huge developer community. The archetype of R is the statistic programing language S[6] and the functional programming language Scheme[7]. R has a big variety of libraries with many different functions for statistical analytics. For graph analysis R provides two common libraries: the Rgraphviz[8] and the igraph[9] library. We are using the later for our implementation because it provides more graph analytics algorithm[10] [26] and it is better applicable for large graphs. The igraph library is also available for other programming languages (e.g., C, Python).

Our graph analytics tool follows a two phases work flow. In the first phase the similarity graph is built and in the second the built graph can be analyzed with different algorithms. The next paragraphs describe the work flow in more detail.

A. Building Similarity Graphs

In the first work flow phase, the similarity graph of RPs of the users of the web community has to be build. We use an undirected, vertices and edges labeled, weighted graph without loop to build a model for the similarity of the web community users. The weighted edges represent the similarity between the vertices which stand for the users. The edges are labeled with the similarity value, that is the Manhattan Distance between the RPs of the users. The labels of the vertices are the user IDs. We use an undirected graph because the similarity of two users can be interpreted in both directions. Figure 2 and 3 show examples of a similarity graph. As mentioned before, in the research field of social networks graph analysis is used to detect social structures between the users, like in [27]. These graphs represent the friend relationship of the users and is in comparison to our work different. We use *weighted* graphs to embody the similarity of users where the edge weights represent the similarity between the interests of the users. So we are not able to use the graph analytics algorithm tools from the social network analysis.

In our tool, the web site owner can chose different alternatives to build a similarity graph for the analysis. The

[5]http://www.r-project.org/

[6]http://stat.bell-labs.com/S/

[7]http://www.r6rs.org/

[8]http://www.bioconductor.org/packages/release/bioc/html/Rgraphviz.html

[9]http://igraph.sourceforge.net/

[10]http://igraph.sourceforge.net/doc/html/index.html

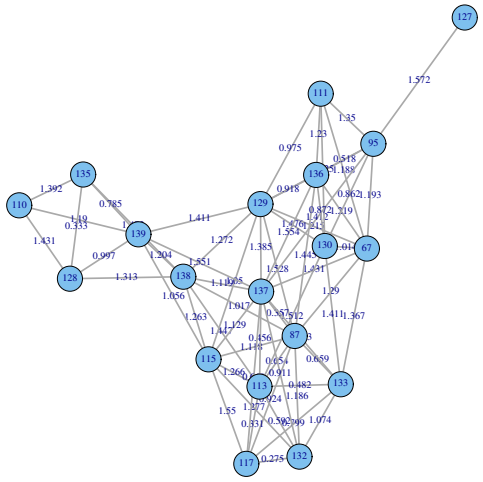


Figure 2. Smallest connection graph.

vertices of the graph (the users) are connected via edges that represent the similarity. It is possible to connect every user to all other users so that a complete graph represents the similarity between all users. This graph is huge and not easy to understand. To reduce the complexity of this graph we introduce a *similarity threshold*. This threshold defines how similar the users must be to be connected together. Only users are connected via vertices whose Manhattan Distance of their RPs is smaller (remember: the smaller the distance the more similar users are) than the chosen threshold. Our analysis tool provides several predefined options to build different graphs with different thresholds. All these graphs are subgraphs of the complete similarity graph of the whole web community:

- **Smallest connected graph**: with this option the threshold increases until every user has at least one connection to another user. In Figure 2, user no. 127 was added last to the graph and has a Manhattan Distance of 1.572. Accordingly all connected vertices have a similarity smaller or equal to 1.572. The result is **one** connected graph.
- **Closest neighbor graphs**: here users are only connected with their most similar neighbors. Every vertex has at least one edge to another vertex. If there exist more most similar neighbors with the same edge weight, the vertex is connected to all of them. This can result in **many** independent graphs as displayed in Figure 3. The difference to the nearest neighbor algorithm is that the nearest neighbor algorithm calculates a path through an existing graph by choosing always the nearest neighbor of the actual vertex.
- **Minimum spanning tree** [28]: is a subgraph where all users are connected together with the most similar users. In contrast to the "closest neighbor graph" we have **one** connected graph.

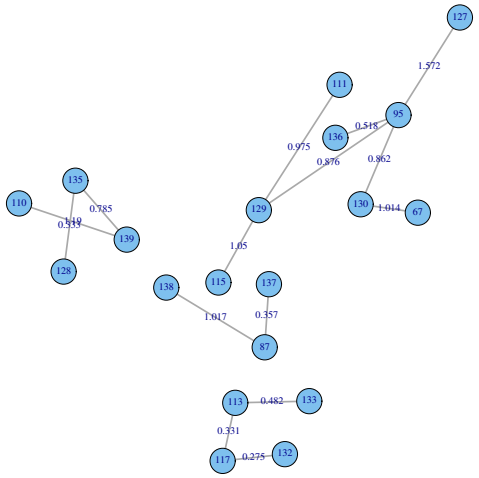


Figure 3. Closest neighbor graphs.

- **Threshold graph**: at last the web site owner can chose a similarity threshold on her/his own. To simplify the choice, the tool suggests two thresholds to the owner: a minimum threshold and a maximum threshold. With the minimum threshold only the most similar users are connected together and with the maximum threshold all users are connected together with every user. So the owner can chose a value between the suggested thresholds to get meaningful results.

B. Similarity Graph Mining Algorithms

In the second work flow phase the web site owner can analyze the graph, generated in the first phase of the work flow, with different algorithms. The aim here is to detect the important users in the graph.

What is an important user? There exists not only one valid definition because it depends–as always–on the point of view. In social networks, e.g., the importance of users often stands for their reputation. The reputation of a user can be measured, e.g., by its number of connectors to other users. Therefore a connector in social networks has another meaning, i.e., the friendship, like in our similarity graphs, we can not use this definition of user importance.

In a social graph a user could be important if she/he is central in respect to the graph. Centrality means that from this very user all other users should be not far away–it should be the nearest neighbor. These highly connected users are often referred as *Hubs* or *Authorities* [24]. Hubs have many outgoing edges while Authorities have many incoming edges.

In a weighted similarity graph high importance could mean that this user is the most similar to other users–she/he should have many edges to other vertices and the edges weights should be as low as possible.

Accordingly, we provide nine algorithms to discover the importance of users. Therefore the importance is defined by the used algorithm which are explained below.

- **PageRank**: The vertex with the highest "PageRank" is the most important user.
- **Jaccard similarity coefficient**: We interpret the most similar vertex as the most important user.
- **Dice similarity coefficient**: Like above we interpret the most similar vertex as the most important user.
- **Nearest neighbors degree**: If a vertex has many neighbors it can be considered as important.
- **Closeness centrality**: Vertices with a low closeness centrality value are important.
- **Hub score**: Vertices with a high score are named hubs and should be important.
- **Eigenvector centrality**: Vertices with a high eigenvector centrality score are considered as important users.

As these seven algorithms above are not *extra* designed to find the important vertices, i.e., users, in similarity graphs of user interests, we developed two new algorithms:

- **Weighted degree**: This simple algorithm choses the vertex with the most connections. Vertices with many connections are important users because they are similar to other user. Actually they are connected with other users cause of their similarity. If there are vertices with the same number of connections it takes the vertex with the lowest edge weights. Therefore the most unimportant vertex has fewer connections to other vertices and the highest edge weights.
- **Range centrality**: The idea behind this algorithm is that a user is important who has many connections in comparison with the other users of the graph, short distance to her/his neighbors, and low edge weights. The range centrality is defined as follows:

$$C_r = \frac{range^2}{aspl + aspw} \tag{7}$$

The *range* is the fraction of the number of users that are reachable from the analyzed vertex and of all users of the graph. We take the square of the range because we consider a user as very important that is connected with many other users:

$$range = \frac{\#reachable\ user}{\#all\ user} \tag{8}$$

The average shortest path length (*aspl*) is the average length of all shortest paths divide by the number of all shortest paths. The shortest paths are calculated with the analyzed vertex as starting point:

$$aspl = \frac{average\ shortest\ paths\ length}{\#shortest\ paths} \tag{9}$$

 ISBN: 978-1-61208-242-4

Table I
EVALUATION RESULTS: IDS OF THE USERS WITH MAXIMUM AND MINIMUM IMPORTANCE OF EVERY GRAPH TYPE (ROWS) FOR DIFFERENT ALGORITHMS (COLUMNS).

		Page Rank	Nearest N.D.	Dice S.C.	Jaccard S.C.	Closeness C.	Hub Score	Eigen-vector C.	Weighted D.	Range C.
SCG	Max	220	222	241	232	93,220	93	106	93	220
	Min	104	138	104	104	64	104	104	104	104
CNG	Max	220	169	79,80,121,200	79,80,121,200	87, 213	66	204	66	87,213
	Min	68	67,127,...	67,127,...	67,127,...	67,127,...	100,246	244	104	67,127
MST	Max	213	169	200	68,80,121	156	261	129	66	156
	Min	104	170	112,126,166	112,126,166	189	189	88	104	189
CG	Max	213	213	all	all	all	all	104	213	213
	Min	104	104	all	all	all	all	241	79	104

With the average shortest path weight (*aspw*) we take into account that the weight of the connected vertices should be very low, i.e., the vertices should be very similar. It's the fraction of the sum of all shortest paths weights and of the number of all shortest paths:

$$aspw = \frac{sum\ of\ all\ shortest\ paths\ weights}{\#shortest\ paths} \quad (10)$$

In the next section we will use our analysis tool with real usage data and compare our new algorithms with the established ones.

V. EVALUATION

A. Material and Methods

To evaluate our algorithms, we use the real usage data from our institute web site[11], i.e., the users' session log files of the site community. We observed 191 registered users over two years. For each user an RP is calculated. Next, we use our analytics tool to build similarity graphs from the RPs of the users and calculate for every graph type the most important and the most unimportant user.

B. Results

Table I displays the results of our calculations. The rows present the different graph types: *SCG* stands for *S*mallest *C*onnection *G*raph, *CNG* for *C*losest *N*eighbor *G*raph, *MST* for *M*inimum *S*panning *T*ree, and *CG* for *C*omplete *G*raph. For every graph type, the user with maximum and minimum importance is displayed. Every column presents one importance algorithm. We can observe the following fact in the dataset in respect to our algorithms, the weighted degree and the rang centrality:

[11] http://www.dbis.cs.uni-frankfurt.de/

In the SCG, the range centrality calculates the same un-/important users like the PageRank and eigenvector closeness, the weighted degree algorithm like the hub score. The majority of algorithms calculate the same unimportant user, only the nearest neighbor degree and closeness centrality differs.

In the CNG, the range centrality and the closeness centrality calculates the same two important users. But the unimportant users are different. The results of the weighted degree for the important user are like the hub score, but the unimportant user is different.

In the MST, the results of the range centrality equals the closeness centrality, while the weighted degree calculates the same unimportant user as the PageRank.

In the CG, user no. 213 is the most important user for both, the weighted degree and the range centrality. The PageRank and the nearest neighbor degree have the same result, only the eigenvector centrality differs. The user no. 104 is the most unimportant user for the rang centrality, the PageRank, and the nearest neighbor degree. The Dice similarity coefficient, Jaccard similarity coefficient, closeness centrality, and hub score are not able to find an un-/important user in the complete graph, because these algorithm do not include the edge weights into their calculation.

C. Discussion

Since there is no objective measurement for importance, we compare established algorithm with our approach. Every algorithm calculates importance in a different way, because every algorithm author has another definition of importance. Most of the algorithms are not designed for similarity or even weighted graphs. Therefore a comparison is not easy.

The weighted degree algorithm firstly focuses on the number of connected neighbors and secondly on the weights of the connected edges. The results of the weighted degree algorithm are very different from the results of the other

algorithm, only the hub score seems to be comparable. In contrast to the hub score the weighted degree algorithm is able to find an important user in a complete graph because it considers the edge weights of the connections (if there are users with the same number of connections which is alway the case in a complete graph).

Similarly, the range centrality focuses on the number of connections, but also on the reachability of the user and the path length. In other words, it considers the whole graph. In comparison to the other algorithms the range centrality is very similar to the closeness centrality but the results differs at the complete graph. Here, our range centrality algorithm calculates important and unimportant users, which is similar to the PageRank algorithm, but the closeness centrality can not calculate any similarity. This is an advantage of our algorithm.

In summary, we think that our new algorithms are a good alternative for finding important users in similarity graphs.

VI. Conclusion

With the results of graph analysis we are now able to answer the research questions of Section III:

- Which are the important users of the web community?
  We provide several algorithms (see Section II-C) to calculate the important user(s) of the community. The definition of importance is dependent from the used algorithm. For example, vertices with many low weight connections can be considered as the important users of the community. These users are very similar to the other users, expressed by the low edge weight.
- Which users have similar interests?
  All users are connected via weighted edges. Users with similar interests have connections with low weights. The web site owner can also define which users are connected together by selecting a similarity threshold (see work flow phase one, Section IV-A). As result only similar users are connected via edges.
- How similar are the interests of the users of the community?
  The lower the weight of the edges the more similar are the users of the community. We give the web site owner the possibility to set thresholds to identify quickly the similarity of her/his community (see Section IV-A).
- How is the community structured? Is it a homogeneous community where every user has similar interests or is it heterogeneous?
  The visualized graph of the community will give the web site owner an overview over the structure of the whole community of her/his web portal.

With answers to these questions, a web site owner is now able to start more focused marketing campaigns. To test new contents or features for her/his web site she/he could start with the most similar users, these users can be considered as an archetype for her/his community.

Acknowledgment

We would like to thank Roberto V. Zicari, Natascha Hoebel, Karsten Tolle, Naveed Mushtaq, and Nikolaos Korfiatis of the Gugubarra team, for their valuable support and fruitful discussions.

References

[1] I. Esslimani, A. Brun, and A. Boyer, "Detecting leaders in behavioral networks," in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 281–285. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2010.72

[2] N. Mushtaq, P. Werner, K. Tolle, and R. Zicari, "Building and evaluating non-obvious user profiles for visitors of web sites," in *IEEE Conference on E-Commerce Technology (CEC 2004)*. Los Alamitos, CA, USA: IEEE Computer Society, 2004, pp. 9–15. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/ICECT.2004.1319712

[3] N. Hoebel and R. V. Zicari, "Creating user profiles of web visitors using zones, weights and actions," in *Tenth IEEE Conference On E-Commerce Technology (CEC 2008) And The Fifth Enterprise Computing, E-Commerce And E-Services (EEE 2008)*. Los Alamitos, USA: IEEE Computer Society Press, 2008, pp. 190–197.

[4] B. Weischedel and E. K. R. E. Huizingh, "Website optimization with web metrics: a case study," in *Proceedings of the 8th international conference on Electronic commerce: The new e-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet*, ser. ICEC '06. New York, NY, USA: ACM, 2006, pp. 463–470. [Online]. Available: http://doi.acm.org/10.1145/1151454.1151525

[5] S. Jung, J. L. Herlocker, and J. Webster, "Click data as implicit relevance feedback in web search," *Information Processing and Management: an International Journal*, vol. 43, no. 3, pp. 791–807, 2007.

[6] N. Hoebel, N. Mushtaq, C. Schefels, K. Tolle, and R. V. Zicari, "Introducing zones to a web site: A test based evaluation on semantics, content, and business goals," in *CEC '09: Proceedings of the 2009 IEEE Conference on Commerce and Enterprise Computing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 265–272.

[7] C. Schefels and R. V. Zicari, "A framework analysis for managing feedback of visitors of a web site," *International Journal of Web Information Systems (IJWIS)*, vol. 8, no. 1, pp. 127–150, 2012.

[8] L. Yi and B. Liu, "Web page cleaning for web mining through feature weighting," in *Proceedings of the 18th international joint conference on Artificial intelligence*, ser. IJCAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 43–48. [Online]. Available: http://dl.acm.org/citation.cfm?id=1630659.1630666

 ISBN: 978-1-61208-242-4

[9] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proceedings of the 8th International Conference on Database Theory*, ser. ICDT '01. London, UK: Springer-Verlag, 2001, pp. 420–434. [Online]. Available: http://dl.acm.org/citation.cfm?id=645504.656414

[10] S.-H. Cha, "Comprehensive survey on distance / similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.8446\&rep=rep1\&type=pdf

[11] L. Euler, "Solutio problematis ad geometriam situs pertinentis," *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, vol. 8, pp. 128–140, 1736.

[12] R. J. Wilson, *Introduction to Graph Theory*. Longman, 1979. [Online]. Available: http://books.google.com.ag/books?id=5foZAQAAIAAJ

[13] D. Jungnickel, *Graphen, Netzwerke und Algorithmen (3. Aufl.)*. BI-Wissenschaftsverlag, 1994.

[14] B. Bollobás, *Modern Graph Theory*, ser. Graduate texts in mathematics. Springer, 1998. [Online]. Available: http://books.google.com/books?id=SbZKSZ-1qrwC

[15] M. A. Rodriguez and P. Neubauer, "Constructions from dots and lines," *CoRR*, vol. abs/1006.2361, 2010.

[16] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.

[17] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the Seventh International World Wide Web Conference*, vol. 30, no. 1-7, 1998, pp. 107–117. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016975529800110X

[18] D. Nemirovsky and K. Avrachenkov, "Weighted pagerank: Cluster-related weights," in *TREC*, 2008.

[19] J. M. Pujol, R. Sangüesa, and J. Delgado, "Extracting reputation in multi agent systems by means of social network topology," in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, ser. AAMAS '02. New York, NY, USA: ACM, 2002, pp. 467–474. [Online]. Available: http://doi.acm.org/10.1145/544741.544853

[20] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, February 1912. [Online]. Available: http://www.jstor.org/stable/2427226

[21] G. Csardi, *Network Analysis and Visualization*, 0th ed., http://igraph.sourceforge.net, August 2010, package 'igraph'.

[22] A. Barrat, M. Barthèlemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004. [Online]. Available: http://www.pnas.org/content/101/11/3747.abstract

[23] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[24] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999. [Online]. Available: http://doi.acm.org/10.1145/324133.324140

[25] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, March 1987.

[26] J. Marcus, "Rgraphviz," Presentation, 2011, accessed: November 4th 2011. [Online]. Available: http://files.meetup.com/1781511/RgraphViz.ppt

[27] L. A. Adamic and E. Adar, "Friends and Neighbors on the Web," *Elsevier Social Networks*, vol. 25, no. 3, pp. 211–230, July 2003.

[28] R. C. Prim, "Shortest connection networks and some generalizations." *Bell Systems Technical Journal*, pp. 1389–1401, November 1957. [Online]. Available: http://www.alcatel-lucent.com/bstj/vol36-1957/articles/bstj36-6-1389.pdf

 ISBN: 978-1-61208-242-4

Characterization of Network Traffic Data

A Data Preprocessing and Data Mining Application

Esra Kahya-Özyirmidokuz, Ali Gezer, Cebrail Çiflikli

Kayseri Vocational College, Erciyes University, Kayseri, Turkey

e-mails: {esrakahya, aligezer , cebrailc} @erciyes.edu.tr

Abstract— **Large amount of traffic data are transmitted during day-to-day operation of wide area networks. Due to the increment of diversity in network applications, its traffic features have substantially changed. Data complexity and its diversity have been rapidly expanding with the changing nature of network applications. In addition, bandwith and speed of network have increased rapidly as compared to the past. Therefore, it is a necessity to characterize the changing network traffic data to understand network behavior. The aim of this research is to understand the data nature and to find useful and interesting knowledge from the network traffic traces which contains IP protocol packets. We analyze the traffic trace of 21 April 2012 on a 150 Mbps transpacific link between US and Japan from the MAWI Working Group traffic archive. This data contain lots of useful and important information which is hidden and not directly accessible. In this research, firstly, anomaly detection analysis and Kohonen Networks are applied to reduce the data matrix. Then, we generate a CART decision tree model to mine traffic data. The decision tree method is successfully applied in network traffic analysis. The results show that the proposed method has substantially good performance.**

Keywords-Network traffic data analysis; Kohonen networks; Data mining; CART; Preporcessing process.

I. INTRODUCTION

Network applications have substantially differed in recent years. Previously, server-client application traffic constitutes most of the Internet traffic. Nowadays, peer-to-peer protocols and applications take a large amount of the total bandwith on the Internet [2]. Also, due to the continuous growth of network speed and bandwith, large amount of traffic are transmitted during day-to-day operation of wide area networks. Therefore, data complexity and its diversity have been rapidly expanding as the increased amount of traffic and changing nature of applications. The characterization of the network traffic data becomes a necessity to understand network behavior. Our objective is using data mining techniques, specifically decision trees, to understand the data characterization and to find useful and interesting knowledge from the network traffic which contain IP protocol packets. We analyzed the traffic trace of 21 April 2012 taken from MAWI Working Group traffic archive. The trace was captured on a 150 Mbps transpacific link between US and Japan. Capturing starts at 2 p.m. and finishes at 2:15 p.m. Due to the large amount of captured packets, we use only 1226014 packets in our analysis. This data contain lots of useful and important information which is hidden and not directly accessible.

After building the database, exploratory data analysis is performed. An approach is presented for pre-processing the data for improving the quality of data and removal of noisy, erroneous and incomplete data. Believability of the data is controlled. Incomplete and inconsistent data analysis are applied. Moreover, anomaly detection analysis is used to reduce the records of data matrix in the data preprocessing process. Although there are many different techniques which can be used in this study, e.g., PCA [20][25], factor analysis [10] and attribute relevance analaysis [3][4], we used Kohonen Networks (KN) in clustering due to the strength of Kohonen maps that lies in their ability to model non-linear relationships between data. In addition, factor analysis, in its forms (PCA, CA, MCA) is the ideal method for providing an overview of the data and continuous spatial visualization of the individuals, and even, in some cases, detecting the natural number of clusters [23]. Kohonen Maps are useful tools for the data mining (DM) models with large data sets. High-dimensional data is projected to a lower dimension representation scheme that can be easily understood. Besides, Kohonen Maps can be used to process qualitative variables as well as quantitative ones [23]. In addition, it can be preferred because of the amount of data matrix. In this research, KN is applied to reduce the database.

There are numerous advantages of hierarchical classifiers based on DTs [23]. We have generated a CART (Classification and Regression Trees) DT (Decision Tree) model in SPSS Clementine 10.1. Colasoft Capsa 6.0 packet sniffer application is used to filter IPV4 packets.

The rest of the paper is organized as follows. The next section presents the literature. In Section III, we give a formal definition of KNs. We describe CART DT in Section IV. Section V presents the application. We draw our conclusion in Section VI.

II. LITERATURE REVIEW

Clustering algorithms have been, and continue to be, widely used for network traffic data classification [11][21][26]. Eshghi et al. [8] compared tree cluster techniques: traditional clustering methods, Kohonen Maps and latent class models. Each methodology could lead to potentially different interpretations of the underlying structure of the data.

There have been many applications of KNs to DM. Larose [6] uses the cluster memberships in data mining. A CART DT model was run, to classify customers, as either churners or nonchurners. Malone et al. [12] demonstrated a trained SOM (Self-Organising Map) which could provide initial information for extracting rules that described cluster

 ISBN: 978-1-61208-242-4

boundaries. They used iris, monks and lungCancer data. Gomez-Carracedo et al. [17] applied Kohonen SOMs to perform pattern recognition in four datasets of roadside soils samples obtained in four sampling seasons along a year. They used CART as an objective variable selection step before the SOM grouping. Siriporn and Benjawan [18] represented an unsupervised clustering algorithm namely sIB, RandomFlatClustering, FarthestFirst, and FilteredClusterer that previously not been used for network traffic classification. Exported network traffic data can be used for a variety of purposes, including DM modeling. Tan et al. [22] conducted network packets analysis on application layer, analyzed network resource location and data size accurately, and studied the traffic characters of network redundancy. Palomo et al. [7] reported that the SOM which was successful for the analysis of highly dimensional input data in DM applications as well as for data visualisation in a more intuitive and understandable manner, had some problems related to its static topology and its inability to represent hierarchical relationships in the input data. To overcome these limitations, they generated a hierarchical architecture GHSOM that is automatically determined according to the input data and reflects the inherent hierarchical relationships among them. Apiletti et al. [5] designed the NETMINE framework which allowed the characterization of traffic data by means of DM techniques.Zaki and Sobh [14] presented an approach to observe network characteristics based on DM framework. They designed a database system and implemented it for monitoring the network traffic.

III. KOHONEN NETWORKS

KNs represent a type of SOM, which itself represents a special class of neural networks. The goal of awl-organizing maps is to convert a complex high-dimensional input signal into a simpler low-dimensional discrete map. Thus, SOMs are nicely appropriate for cluster analysis, where underlying hidden patterns among records and fields are sought. A typical SOM architecture is shown in Figure 1 [6].

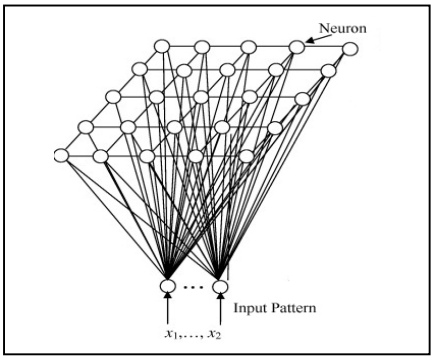


Figure 1. Topology of a simple SOM.

KNs can be considered a non-hierarchical method of cluster analysis. As non-hierarchical methods of clustering, they assign an input vector to the nearest cluster, on the basis of a predetermined distance function but they try to preserve a degree of dependence among the clusters by introducing a distance between them. Consequently, each output neuron has its own neighborhood, expressed in terms of a distance matrix. The output neurons are characterized by a distance function between them, described using the configuration of the nodes in a unidimensional or bidimensional space [19].

KNs perform unsupervised learning; an *Out* field is not specified and the model is, therefore, not given an existing field in the data to predict. KNs attempt to find relationships and overall structure in the data. The output from a Kohonen network is a set of (X, Y) coordinates, which can be used to visualize groupings of records and can be combined to create a cluster membership code. It is hoped that the cluster groups or segments are distinct from one another and contain records that are similar in some respect [24].

Suppose that we consider the set of *m* field values fot the *n*th record to be an input vector $x_n = x_{n1}, x_{n2}, \ldots x_{nm}$, and the current set of m weights for a particular output node *j* to be a weight vector $w_j = w_{1j,} w_{2j,\ldots} w_{mj}$. In Kohonen learning, the nodes in the neighborhood of the winning node adjust their weights using a linear combination of the input vector and the current weight vector:

$$w_{ij,new} = w_{ij,current} + \eta(x_{ni} - w_{ij,current}) \qquad (1)$$

where η, $0 < \eta < 1$, represents the learning rate, analogous to the neural networks case. Kohonen indicates the learning rate should be a decreasing function of training epochs (runs through the data set) and that a linearly or geometrically decreasing η is satisfactory for most purposes. The algorithm of KNs is shown in the accompanying box [6].

For each input vector *x*, do:

- Competition. For each output node j, calculate the value $D(w_j, x_n)$ of scoring function. For example, for Euclidean distance, $D(w_j, x_n) = \sqrt{\sum_i (w_{ij}, x_{ni})^2}$. Find the winning node *j* that minimizes $D(w_j, x_n)$ over all output nodes.
- Cooperation. Identify all output nodes j within the neighborhood of J defined by the neighborhood size R. For these nodes, do the following for all input record fields:
- Adaptation. Adjust weights: $w_{ij,new} = w_{ij,current} + \eta(x_{ni} - w_{ij,current})$
- Adjust the learning rate and neightborhood size, as needed.
- Stop when the termination criteria are met.

IV. DECISION TREES

The DT technique is one of the most intuitive and popular DM methods, especially as it provides explicit rules for classification and copes well with heterogeneous data missing, data and non-linear effects [23]. There are numerous advantages of hierarchical classifiers based on DTs. DTs provide an easy to understand overview for users without a DM background with high classification

accuracy. They also provide a tree model of the problem and various alternatives in an understandable format without explanation. The acquired knowledge are usually quite understandable and can be easily used to obtain a better understanding of the problem. In addition, DTs assist in making decisions with existing information. They have satisfactory performance even when the training data is highly uncertain.

A. CART Decision Tree

CART is a popular DT algorithm first published by L. Breiman, J. Friedman, R. Olshen, C. Stones in 1984 [13]. The CART algorithm grows binary trees and continues splitting as long as new splits can be found that increase purity. The CART algorithm identifies a set of such sub-trees as candidate models. These candidate sub-trees are applied to the validation set, and the tree with the lowest validation set misclassification rate (or average squared error for a numeric target) is selected as the final model [16].

The CART algorithm identifies candidate sub-trees through a process of repeated pruning. The goal is to prune first those branches providing the least additional predictive power per leaf. To identify these least useful branches, CART relies on a concept called the adjusted error rate. This is a measure that increases each node's misclassification rate or mean squared error on the training set, by imposing a complexity penalty based on the number of leaves in the tree. The adjusted error is used to identify weak branches (those whose error enough to overcome the penalty) and mark them for pruning. The formula for adjusted error rate is

$$AE(T) = E(T) + \alpha leaf\_count(T) \quad (2)$$

where α is an adjustment factor that is increased in gradual steps to create new subtrees. When α is 0, the adjusted error rate equals the error rate. The algorithm continues to find trees by adjusting α and pruning back one node at a time, creating a sequence of trees, α_1, α_2, and so on, each with fewer and fewer leaves. The process ends when the tree has been pruned all the way down to the root node. Each of the resulting subtrees (sometimes called alphas) is a candidate to be the final model. Notice that all the candidates contain the root node and the largest candidate is the entire tree [16].

The next step is to select, from the pool of candidate sub-trees, the one that works best on new data. That, of course, is the purpose of the validation set. Each of the candidate sub-trees is used to classify the records or estimate values in the validation set. The tree that performs this task with the lowest overall error is declared the winner. The winning sub-tree has been pruned sufficiently to remove the effects of overtraining, but not so much as to lose valuable information [16]. The winning sub-tree is selected on the basis of its overall error when applied to the validation set. But, while one expects that the selected sub-tree will continue to be the best model when applied to other data sets, the error rate that caused it to be selected may slightly overstate its effectiveness. There may be many sub-trees that all perform about as well as one selected. To a certain extent, the one of these that delivered the lowest error rate on the validation set may simply have "gotten lucky" with that particular collection of records. The selected sub-tree is applied to a third preclassified data set, the test set. The error obtained on the test set is used to predict expected performance of the model when applied to unclassified data [16].

CART uses the Gini index. The Gini impurity is,

$$I_E(m) = -\sum_{i=1}^{k(m)} \pi_i \log \pi_i \quad (3)$$

where π_i are the fitted probabilities of the levels present in node m, which are at most *k(m)* [19].

V. MODELING

This study uses the DM methodology CRISP-DM (Cross-Industry Standard Process for Data Mining) which is represented in Figure 2. According to CRISP-DM, a given DM project has a life cycle consisting of six phases as illustrated in Figure 2 [6].

A. Data Understanding

Preprocessing is an important step for successful DM to analyze the datasets before DM process starts. This section is focused namely on the second and third steps of Figure 2: data understanding and data preparation. The first step in the process is to transfer the data in a database, and make use of a statistical analysis tool to get the details of the network traffic data.

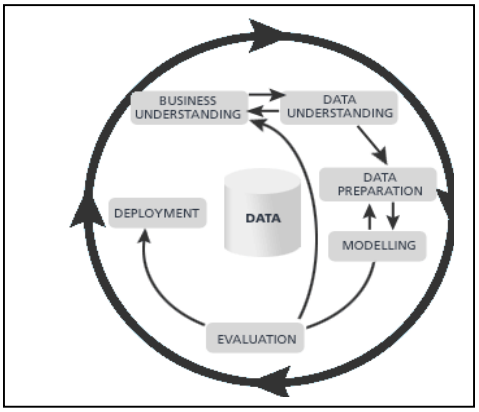


Figure 2. CRISP-DM is an iterative, adaptive process.

A large amount of data is used in this research. The trace data are obtained from MAWI traffic trace archieve. The trace data are captured on a 150 Mbps backbone line that connects US and Japan. Nearly 35 million packets and 31800000000 bytes are captured in 15 minutes time. In this analysis, we only use 1226014 packets captured on 21 April 2012. To filter IP protocol Colasoft Capsa 6.0 packet sniffer application is used.

A flow is identified as the combination of the following attributes which are presented and used in the data set are

{No, Date, Absolute time, Delta time, Relative time, Source, Destination, Protocol type, Size, IP identifier, Source physical, Destination physical, Source IP, Destination IP, Source port, Destination port, Comment, Summary}.

We eliminate date, comment, and summary attributes because they are useless and redundant in the model. The frequencies of protocol attribute variables are seen from Figure 3. In the preprocessing process, protocol attribute variables which are less than 0,01% are combined as 'others' variable. These updated variables are: MSSQL, POP3, RTCP, H.225, IMAP, BGP, QQ, SIP, POP3s, PPTP, MGCP, NFS, Telnet, H.323, PIM, LPD, SAP, WINS, IP, LDAP, PDM, NBDGM, NBSSN, SLP and Whols. SQL codes are used in this variable reduction. In the data matrix, there are 329 records whose protocol attribute variables are 'others'.

B. Anomaly Detection

DM uses huge amounts of data with many thousands or even millions of records, outliers and unusual data should be explored when preparing data for modeling. Anomalous data is a problem for models. In this research, anomaly detection analysis is used to find data values which show different behavior from the previous measured values. Anomaly detection is an important data mining task. It should be done in the data preprocessing process.

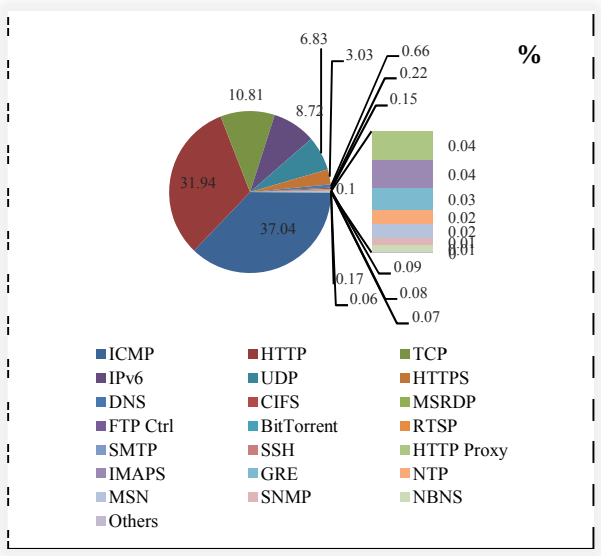


Figure 3. Protocol Attribute Variables

Anomaly detection models are used to identify outliers, or unusual cases in the data. Anomaly detection can examine large numbers of fields to identify clusters or peer groups into which similar records fall. Each record can then be compared to others in its peer group to identify possible anomalies. After anomaly detection analysis, an anomaly index is calculated for each record which is the ratio of the group deviation index to its average over the cluster that the case belongs to [3]. In this study, anomaly detection is used to locate the records that are the most unusual with respect to those fields. 12260 anomalous records are identified with eight clusters; that is the 1% of the cases that we requested. Average anomaly index level is 3.14041. Records which are greater than the average anomaly index level are selected as anomaly records. For example, record-1's protocol attribute value is "http". After anomaly detection, the record-1 which has "http" protocol value is determined in group 4. This record's anomaly index level is 1.079. We cannot eliminate the record because of its low index. 12260 of records are eliminated using the anomaly detection algorithm with the SPSS Clementine 10.1 DM Tool.

C. Data reduction via Kohonen Networks

Data reduction is an important stage for data preprocessing. The reduction technique is very important because as the data space is reduced, information can be lost. The reduced data must give a complete picture of the data to the analyst.

Cluster membership may be used to enrich the data set and improve model efficacy. Indeed, as data repositories continue to grow and the number of fields continues to increase, clustering has become a common method of dimension reduction [6].

KNs require us to specify the number of rows and the number of columns in the grid space characterizing the map. Large maps are usually the best choice, as long as each cluster has significant number of observations. The learning time increases significantly with the size of the map. The number of rows and the number of columns are usually established by conducting several trials until a satisfactory result is obtained [19].

The Kohonen SOM has several important properties that can be used within the DM/knowledge discovery and exploratory data analysis process. A key characteristic of the SOM is its topology preserving ability to map a multi-dimensional input into a two-dimensional form. This feature is used for classification and clustering of data. [12]. Standard clustering methods do not handle truly large data sets well, and fail to take into account multi-level data structures [1]. In this study, we use KNs to reduce the attributes of the clustering model. Therefore, only selected attributes are then kept to represent the document collection, the remaining ones are discarded.

Records are grouped by KN so that records within a group or cluster tend to be similar to each other, and records in different groups are dissimilar. The Kohonen parameters were set in SPSS Clementine 10.1 as follows. For the first 20 cycles, the neighbourhood size was set at R=2, and the learning rate was set to decay linearly starting at $\eta = 0.3$. Then, for the next 150 cycles, the neighborhood size was reset to R=1 while the learning rate was allowed to decay linearly from $\eta = 0.3$ to $\eta = 0$.

The neurons are organized into two layers, input layer and output layer. In the study, the input layer has 6 neurons and the output layer has 70 neurons.

 ISBN: 978-1-61208-242-4

The most attractive feature of the Kohonen SOMs is that, once trained, the map represents the projection of the data set belonging to an N-dimensional space into a bi-dimensional one [9]. The KN performs the projection of the H–dimensional space, containing the X vectors representing the load diagrams, into a bi-dimensional space. When a data stream passes through the generated model Kohonen node, two new fields are created, representing X- and Y-coordinates of the clusters. The clusters are identified in the Kohonen output window by their values on these coordinates [24]. Two coordinates $KX(10) and $KY(7) are representing the Kohonen net attributes.

A nine by six Kohonen clustering is performed. The unwanted attribute absolute time, destination and source attributes are blocked, leaving only 8 attributes. Delta time, relative time, IP identifier, destination physical, source physical, and size attributes and Kohonen-X and Kohonen-Y are selected as important attributes. The elimination of absolute time is expected because absolute time attribute has a direct relationship with delta time and relative time attributes.

D. Data mining modeling

There are many different types of classification methods. The choice of the best predicting technique depends on the data set being analyzed, and its complexity, the time it takes to generate, and the results of the analysis.The output attribute is discrete. In addition, DTs are powerful and popular tools for classification and prediction.

In this section, we present data mining modeling phase which is the fourth step of Figure 2. purpose of the decision tree is to provide a simple and understandable model of data. A CART DT model is run to classify network traffic data to build a DT for predicting network protocol type. The aim of the classification is to find the similar data items which belong to the same class.

Gini index is used. The depth of the tree is limited to 5 levels below the root node. The stopping criteria details are as follows: the minimum records in parent branches are 4%, and the minimum records in child branches are 2%. The minimum change in impurity is 0,0001. Maximum surrogates are 5.

In the CART model, the target is the protocol attribute. Delta relative size, IP identifier, source physical and destination physical are the input attributes. The model is given in Figure 4. We can show the tree model with if-then rules to express the process in English. 13 rules are generated. The following examples illustrate some of the rules:

- If [ip identifier=<46.5] and [size<=1.499] then [protocol is ipv6].
- If [ip identifier>46.5] and [size>64.5] and [size<=70.5] then [protocol is TCP].

E. Accuracy

The cross-validation method involves partitioning the examples randomly into *n* folds. (Ten is a fairly popular choice for *n*, but much depends on the number of examples available). We use one partition as a testing set and use the remaining partitions to form a training set. As before, we apply an algorithm to the training set and evaluate the resulting model on the testing set, calculating the percentage correctly. We repeat this process by using each of the partitions as the testing set and using the remaining partitions to form a training set. The overall accuracy is the accuracy averaged over the number of runs, which is equivalent to the number of partitions. *Stratified* cross-validation involves creating partitions so that the number of examples of each class is proportional to the number in the original set of examples [15].

```
IP Identifier <= 46,500 [ Mode: ICMP ]
    Size <= 74,500 [ Mode: ICMP ] => ICMP
    Size > 74,500 [ Mode: IPv6 ] => IPv6
IP Identifier > 46,500 [ Mode: HTTP ]
    Size <= 32,759 [ Mode: HTTP ]
        Size <= 1,508 [ Mode: HTTP ]
            Size <= 1,498 [ Mode: HTTP ]
                Size <= 1,394 [ Mode: UDP ] => UDP
                Size > 1,394 [ Mode: HTTP ] => HTTP
            Size > 1,498 [ Mode: IPv6 ] => IPv6
        Size > 1,508 [ Mode: HTTP ]
            Source Physical in [ "00:0E:39:E3:34:00" ] [ Mode: HTTP ]
                IP Identifier <= 24.484 [ Mode: HTTP ] =>
                HTTP
                IP Identifier > 24.484 [ Mode: HTTP ] =>
                HTTP
            Source Physical in [ "00:16:9C:7C:B0:00" ] [ Mode: HTTP ]
                => HTTP
    Size > 32,759 [ Mode: ICMP ]
        Source Physical in [ "00:0E:39:E3:34:00" ] [ Mode: HTTP ]
            Size <= 70,500 [ Mode: HTTP ] => HTTP
            Size > 70,500 [ Mode: UDP ]
                Size <= 188,500 [ Mode: UDP ] => UDP
                Size > 188,500 [ Mode: HTTP ] => HTTP
        Source Physical in [ "00:16:9C:7C:B0:00" ] [ Mode: ICMP ]
            Size <= 64,500 [ Mode: ICMP ]
                IP Identifier <= 30602,500 [ Mode: ICMP ] =>
                ICMP
                IP Identifier > 30602,500 [ Mode: ICMP ] =>
                ICMP
            Size > 64,500 [ Mode: TCP ]
                Size <= 70,500 [ Mode: TCP ] => TCP
                Size > 70,500 [ Mode: ICMP ] => ICMP
```

Figure 4. Data Mining Model

Tenfold cross-validation accuracy evaluation is used to train and test the data matrix. Records (917984) of data (1213754) are correct. The accuracy ratio of the model is 75.63% which is shown in Table III.

TABLE I. MODEL ACCURACY

	Number	Ratio
Correct	917984	75,63%
Incorrect	295770	24,37%

VI. CONCLUSION AND FUTURE WORK

DM is finding wide application in many fields. Network traffic characterization is one of them. Due to the increasing

diversity of network applications, their packet features substantially have changed. Also the growth in network speeds and bandwidths increases the amount of traffic on networks. Therefore, it is a necessity to characterize new traffic features to handle network problems and to increase performance.

In this study, we presented knowledge induction from network traffic data captured on a 150 Mbps trans-Pacific line between Japan and US. Preprocessing techniques were used to improve the quality of data. Noisy, erroneous, and incomplete data were removed from the data matrix. Anomaly detection analysis was used to reduce the data matrix. Moreover, this research included attribute reduction using KNs. A CART DT which uses for classification of the data set with five tree depths is generated. The accuracy ratio of the model is 75.63%.

In this research, network traffic data are mined and useful relationships, groupings, associations are discovered. The DT model lay out the problem clearly so that all options can be explored. This acquired knowledge will be used to predict the future behaviors of the line. In addition, the DT model helps network operators to understand the behavior of network users. This research is also important to assess future network capacity requirements and to plan future network developments.

For the future we plan the further evaluation and implementation of this framework.

REFERENCES

[1] A. Ciampi and Y. Lechevallier, "Clustering large, multi-level data sets: an approach based on Kohonen self-organizing maps", Principles of Data Mining and Knowledge Discovery 4th European Conference PKDD 2000 Proceedings Lecture Notes in Artificial Intelligence, vol. 1910, Springer-Verlag, Berlin, Germany, pp. 353-358, 2000.

[2] C. Çiflikli, A. Gezer, A.T. Özşahin, and Ö. Özkasap, "BitTorrent packet traffic features over IPv6 and IPv4", Simulation Practice and Theory, vol. 18, iss. 9, October 2010, pp. 1214-1224.

[3] C. Ciflikli and E. Kahya-Özyirmidokuz, "Enhancing product quality of a process", Industrial Management and Data Systems, vol. 112, iss.8, pp. 1181-1200, 2012.

[4] C. Ciflikli and E. Kahya-Özyirmidokuz, "Implementing a data mining solution for enhancing carpet manufacturing productivity", Knowledge-Based Systems, vol. 23, 2010, pp. 783–788.

[5] D. Apiletti, E. Baralis, T. Cerquitelli and V. D'Elia, "Characterizing network traffic by means of the NETMINE framework", Computer Networks, vol. 53, pp. 774–789, 2009.

[6] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, NJ, USA: Wiley. 2005.

[7] E. J. Palomoa, J. North, D. Elizondo, R.M. Luquea and T. Watson, "Application of growing hierarchical SOM for visualisation of network forensics traffic data", Neural Networks, vol. 32, pp. 275-284, 2012.

[8] Eshghi, D. Haughton, P. Regrand, M. Skaletsky, and S. Woolford, "Identifying groups: A comparison of methodologies", Journal of Data Science, vol. 9, pp. 271-291, 2011.

[9] F. Rodrigues, J. Duarte, V. Figueriredo, Z. Vale, and M. Cordeiro, "A comparative analysis of clustering algorithms applied to load profiling", Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, vol. 2734, pp. 73-85, 2003.

[10] H. F. Wang and C. Y. Kuo, "Factor analysis in data mining", Computers & Mathematics with Applications, vol. 48, iss: 10-11, pp. 1765-1778, 2004.

[11] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms, in: MineNet '06, ACM Press, New York, NY, USA, 2006, pp. 281–286.

[12] J. Malone, K. McGarry, S. Wermter, and C. Bowerman, "Data mining using rule extraction from Kohonen self-organising maps", Neural Comput& Applic, vo.15, pp. 9-17, 2005.

[13] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and Regression Trees, Chapman and Hall/CRC. 1984.

[14] M. Zakia and T. S. Sobhb, "NCDS: Data mining for discovering interesting network characteristics", Information and Software Technology, vol. 47, pp. 189–198, 2005.

[15] M. A. Maloof, Machine Learning and Data Mining for Computer Security, Methods and Applications. USA: Springer-Verlag. 2006.

[16] M. J. A. Berry and G.S. Linoff, Data Mining Techniques for Marketing, Sales, and Customer Relationship Management, Third Ed.. NY: Wiley. 2011.

[17] M. P. Gomez-Carracedo, J.M. Andrade, G.V.S.M. Carrera, J. Aires-de-Sousa, A.Carlosena, and D. Prada, "Combining Kohonen neural networks and variable selection by classification trees to cluster road soil samples", Chemometrics and Intelligent Laboratory Systems, vol. 102, pp. 20-34, 2010.

[18] O. Siriporn and S. Benjawan, "Anomaly detection and characterization to classify traffic anomalies, Case Study: TOT Public Company Limited Network", World Academy of Science, Engineering and Technology, vol. 48, pp. 407-415, 2008.

[19] P. Giudici, Applied data mining. England: Wiley, 2003.

[20] Q. Guo, W. Wu, D.L. Massart, C. Boucon and S. de Jong, "Feature selection in principal component analysis of analytical data", Chemometrics and Intelligent Laboratory Systems, vol. 61, pp. 123-132, 2002.

[21] Q. Wang, V. Megalooikonomu, A clustering algorithm for intrusion detection, Proceedings of SPIE 5812, pp. 31–38, 2005.

[22] S. Tan, M. Chen, G. Yang, and Y. Wang, "Research on Network Data Mining Techniques", Energy Procedia, Singapore, vol. 13, pp. 4853 – 4860, 2011 [ESEP 2011, 9-10 December 2011].

[23] S. Tuffer, Data Mining and Statistics for Decision Making. Wiley, 2011.

[24] SPSS Inc., Introduction to Clementine and Data Mining, Chicago, 2003, http://homepage.univie.ac.at/marcus.hudec/Lehre/WS%202006/Methoden%20DA/IntroClem.pdf [retrieved: 08, 2012].

[25] W. Melssen, R. Wehrens and L. Buydens, "Supervised Kohonen networks for classification problems", Chemometrics and Intelligent Laboratory Systems, vol. 83, pp. 99–113, 2006.

[26] Y. Guan, A. Ghorbani, and N. Belacel, "Y-Means: a clustering method for intrusion detection", Proceedings of Canadian Conference on Electrical and Computer Engineering, 2003, pp. 4–7.

 ISBN: 978-1-61208-242-4

Structured Data and Source Code Analysis for Financial Fraud Investigations

Joe Sremack
Financial and Enterprise Data Analytics
FTI Consulting
Washington, DC USA
joseph.sremack@fticonsulting.com

Abstract— **Financial fraud investigations are becoming increasingly complex. The volume of data continues to increase along with the volume and complexity of underlying source code logic. As the volume and complexity increase, so too does the importance of identifying techniques for reducing the data to manageable sizes and identifying fraudulent activity as quickly as possible. This paper presents how to ensure that all data was properly collected and a methodology for reducing the complexity of such investigations by identifying similarities and differences between the source code and structured data.**

Keywords-structured data analysis, source code review, fraud investigation.

I. INTRODUCTION

Structured data and proprietary source code are two of the most critical sources of information for large-scale financial fraud investigations. Proprietary source code is used to execute in-house investment strategies for investment banks, hedge funds, and other financial institutions. In the financial setting, source code review yields information about how the organization carried out its operations. The second source of information, structured data, are an organized form of data in which connected data are stored in a discrete, atomic form. Structured data are continuously generated during the course of business, and most business events and transactions create structured data that chronicle the organization's history – most notably the financial transactions. The most common form of structured data is data stored in databases. Together, structured data analysis and source code review can reveal how a business operated in a way that is not possible with only one method.

Financial fraud investigations introduce unknowns about the quality and completeness of both the source code and structured data that were produced. Since most transactions are generated from automated events from the source code, analyzing both in junction with each other is critical for not only validating the completeness of both, but also how the organization truly operated. Falsifying both the transactions and the source code together so that there are no discrepancies is infeasible for virtually any type of fraud. The complexity of synchronizing the source code and the structured data, while carrying out the fraud and creating falsified financial reports, is beyond the capabilities of even the best fraud operatives.

The complexity of synchronizing source code and structured data is what makes analyzing both together so critical. Several failed attempts at uncovering financial frauds have demonstrated that merely analyzing the transactions is not enough, the most famous of which is the Bernard Madoff Ponzi scheme scandal [1]. Analyzing structured data alone does not necessarily tell you how the data entered the data repository or what data was excluded, modified, or code-generated. Likewise, analyzing the source code alone does not provide sufficient evidence, since the source code does not necessarily contain information on what steps were actually run and when, nor the extent of the fraud. The source code would most likely have parameters and input data passed into it, and the data could be altered outside of the source code environment. Combining source code review with the structured data analysis identifies data points and values that could not be generated from a normal, non-fraudulent course of business, such as account values and financial charges.

Analyzing source code in a fraud investigation setting is a complex and time-consuming process. A fraud investigation typically hinges on identifying key anomalous transactions or data patterns that diverge from normal business operations and then identifying how the fraud was conducted within the transactions, source code, and business processes. Most key employees have either been laid off or fired when a financial firm is accused of fraud. As such, information about the source code and locating key documentation is difficult or impossible. The source code may be poorly documented, which requires identifying which files and sections of code need to be reviewed, and volume may be in the tens of thousands of files and millions of lines of code. Reviewing every line of code would not be realistic, regardless of the number of analysts. The culling process of reducing the amount of code that needs to be reviewed requires a precise process that reduces the volume to a manageable size for review but does not exclude key information.

This paper discusses the methodology for performing source code review and structured data analysis that have been applied in several financial fraud investigations. Elements of this methodology have been employed on several large-scale financial fraud investigations. The second section covers the general observations an investigator looks for during the course of this type of

 ISBN: 978-1-61208-242-4

investigation. The next section covers the basics of collecting and validating both sources of information. The fourth section discusses data element mapping and function call mapping, which is followed by a section on data value mapping. The concluding section summarizes the process.

II. CURRENT RESEARCH

Current research in source code analysis and classification have yielded several techniques that work well under ideal conditions, but those techniques are not always well-suited to real-world fraud investigations. Major research has been conducted on creating dependence graphs and semantic analysis [2][3]. Most of the topical topics and challenges related to source code analysis are based on those outlined in the seminal paper "Reverse Engineering: A Roadmap." Several techniques, such as island parsing and lake parsing, are better suited for the constraints of a large-scale fraud investigation [5]. Likewise, the field of structured data reverse engineering has produced techniques and a roadmap for analysts [6][7]; however, the need to understand the relationship between the source code and the structured data has not been addressed for practical, complex investigations.

A modern approach to reverse engineering source code is the use of Unified Modeling Language (UML) tools to automatically identify and document source code language constructs, call maps, program behavior, and architecture [8][9]. A common standard that is based on UML has developed called the Knowledge Discovery Metamodel (KDM). The model is based on an Object Management Group Standard that has become an ISO standard in 2012 [10]. Practical objections to UML-based reverse engineering approaches, including KDM, limit the usefulness of using such an approach when time limitations exist and prior knowledge of the relationships of the source code is required [11]. For purposes of a fraud investigation, the total set of source code need not be analyzed, and the analysis should assist with limiting the amount of information that needs to be analyzed. Moreover, the time required for setting up a KDM or other UML-related documentation process with an unknown set of source code can be more time-consuming than operating without any such tool.

Fraud detection and reverse engineering research is a growing field because of the proliferation of cases of fraud and the increasingly complex manner in which they are conducted. The majority of current financial fraud detection literature is based on the analysis of financial transactions using anomaly detection data mining approaches (e.g., Bayesian belief networks, neural networks, and cluster analysis) [12][13][14][15]. The assumption throughout the majority of the literature is that data is pre-cleansed and do not incorporate information about the systems that generated the data.

III. ANALYSIS MOTIVATIONS

There are several purposes for performing this type of analysis. One purpose is to identify how the source code and structured data relate to one another and whether they both tell the same story about the business operations that were performed. The relationship has many dimensions and attributes and depends on the layout of the structured data and the function(s) of the source code. The relationship depends on how the source code affects the structured data and whether the structured data fully adheres to the rules in the source code. For example, one set of source code can modify all sales transaction fields except for customer service inquiries, so the relationship is defined on those fields based on how the structured data adheres to the rules in the source code. The structured data may vary from the source code rules, and that difference shows where other means for modifying the data exist.

A second purpose is to identify key data points in the structured data. Some structured data contains cryptic field names, obscure data values, and a large volume of objects. Analyzing which data points the source code affects can help with defining what certain fields contain. The source code will show where the outputs of the code are stored within the structured data and which inputs and operations are performed on the data before they are stored. Several examples of these are database inserts, deletes, and updates and the generation of output files that are later stored in the database. The process of identifying which inputs and functions generate which outputs in the structured data can also be performed to pare down the number of fields for analysis. The key fields can be identified by locating the outputs from the source code that either have the input sources or the actions that are critical for the investigation. Likewise, some fields may simply be unimportant for the investigation; identifying the unimportant fields that can be ignored in future analyses is valuable for reducing the complexity of the investigation.

Third, the structured data can be compared to the data modification rules in the source code to detect whether any of the data were modified outside of the source code. In the case of a Ponzi scheme, stock trades may be entered by a non-fraudulent program, while a fraudulent process will later correct those trades to represent that different trades were executed. This is critical in fraud investigations because of the possibility of data modifications that occur outside of normal business processes. For example, a rogue program or manual data manipulation can be used to perpetrate the fraud, and the data modified by a rogue process can be difficult to detect otherwise. Several other possibilities, such as a different version of the code having been run or the code was later modified to appear to be non-fraudulent, can exist, but the comparison of source code rules to structured data will detect these differences regardless.

 ISBN: 978-1-61208-242-4

The fourth purpose is to reduce the number of lines of source file code that need to be analyzed. A common problem in any analysis is reducing the volume of data to a manageable size without compromising the completeness of the relevant analysis. The relevant sections in the source code can be more quickly located by analyzing where the key data fields are manipulated and how. A casual chain can be created from those sections to create a network of related sections, and any isolated sections of code can be more easily categorized or deemed non-relevant.

This paper highlights the key types of analysis techniques that have been applied in practical situations with success. As such, these techniques form an agile toolbox for quickly and effectively analyzing large volumes of source code and structured data in the absence of adequate documentation for financial fraud investigations. While the techniques listed are known in the areas of data and source code reverse engineering, their usefulness and practicality have neither been documented in current research circles, nor presented in a manner in which the combined source code and structured data can reside in a single analysis repository.

IV. DATA COLLECTION

The initial steps of any investigation are to collect all data and documentation and verify that the collection is complete. The first step is to survey the data and test that all objects – such as schemas and views – exist in the copied data. Next, comparisons of control totals from the source system to the copied data are performed to provide extra assurance that no data were lost or corrupted. For example, the summation of several numeric columns and the counts of distinct text values in several columns across every table between the source system and copied data are compared and verified. Finally, all available documentation about the source system are collected. The documentation should cover data information (e.g., Entity-Relationship Diagrams, Data Dictionaries, and data value definitions) and business purposes and use (e.g., list of system users and business requirements) [16].

Like structured data analysis, source code review does not begin until after validating the data collection and gathering all supporting documentation. Unlike with structured data analysis, however, ensuring that all source code has been collected is much more complex. Determining if the complete set of source code is available is a critical step that requires reviewing additional documentation, and this step sometimes requires that the source code in question be compiled.

Compiling source code is a dependable method for validating each individual program, for if it does not compile, an issue is known. Compiling another institution's source code is rarely a simple operation, though. Many obstacles make compiling unfeasible, such as compiler settings, compiler versions, and the availability of third-party and custom library files. As a result of these obstacles, compiling the source code is not always possible.

Source code documentation is critical for quickly understanding how the program operates, what functions it serves, the entry points to the source code, and which individual files constitute the complete set of source code. Comments embedded in the source code are valuable, but they rarely provide any insight into the business purpose of the source code or the function call order of the source code. Comments alone also do not offer a reliable method for determining versioning, business purpose, or testing. The following types of documentation are some of the standard documentation that should be collected:

- Compiler logs;
- Source code change history;
- Use case testing documentation;
- Configuration management documentation, and
- Business requirements documentation.

Not all companies fully document all of their source code. Interviews or depositions of programmers and key business owners can greatly assist with understanding how the source code functions and how to analyze the structured data [17]. These interviews, however, cannot always be conducted due to employees refusing to be interviewed or an inability to otherwise question them. These difficulties make the analysis more difficult and are further motivation for why source code and structured data should be analyzed together.

V. ANALYSIS TECHNIQUES

The analysis of source code and structured data is an iterative process of conducting a series of related analyses that assist with the investigation. This section details three analysis methods that have been successfully employed on several large-scale financial fraud investigations. The methods are: data element mapping, function call mapping, and data value analysis. Together, the analysis techniques in this section provide a template for streamlining the analysis process and identifying key relationships between the structured data and the source code. As with any analysis, the methods needed for each investigation vary depending on requirements and available information. Moreover, the techniques in this section do not constitute a full methodology for any type of investigation. Standard analysis techniques for the structured data and source code should still be conducted (e.g., structured data surveying and searching the source code for particular functions, such as file printing).

A. Data Element Mapping

Data element mapping is a process for searching the set of source code by the field and object names in the structured data. The process is often referred to as "crawling code," whereby a code base is scanned for particular values or operations [18]. The information gained from this process is a listing of source code files and line numbers where there

are search term matches on particular data field and object names. Data element mapping is typically conducted before the other two steps, for it offers both a survey of the source code and data elements, as well as a set of potential entry points for the analysis.

The first step is to collect the set of structured data field and object names that are to be searched. All potentially relevant names should be included, such as:

- The database name;
- Schema names;
- Known database object owners;
- Table names;
- XML tags, and
- Table field names.

Common names that would result in too many false positive keyword search matches should be excluded from the search set, including "date," "ID," "comment," and "owner."

Next, the searching is performed against the source code. The source code should be indexed in a document repository, which is an application that allows for keyword searching and regular expression matching, and the keywords should be searched, with the resulting output containing a list of all keyword matches, the file and line where the match occurred, and the matching line of text from the source code.

The output is then reviewed to compile a list of key source code files. The results should be quantified by matches per keyword, and any keyword with a number of results that are anomalously too high or too low should be reviewed further to ensure that they had proper matches. Too many matches can be explained by a keyword that is often used in the source code language or is an alias that has meaning beyond the structured data object name. Too few matches can sometimes be the result of the field or object name not being explicitly called. Further analysis of the other fields in its table or related objects is required to determine if an alias or function is manipulating that field or object instead. For example a stored procedure within a database may be called instead. Any outliers that are identified that are truly anomalies should be removed from the result set.

The resulting set should be quantified once more to generate a count of search matches per source code file. The source code files with the most search matches are to be the starting points for the investigation. These are the files that generate the most transactional activity and should, at a minimum, be reviewed for the structured data operations. In addition, any source code files that had search term matches for the most critical object or field names are critical sources of information that should be analyzed in depth.

B. Function Call Mapping

The second technique is creating a mapping of how the various elements within the source code relate to one another, which is called function call mapping. The purpose is to identify additional sections of the source code that relate to other critical sections. Function call links from the critical sections of source code from data element mapping or other processes are used to locate additional code that may not have been initially deemed to be relevant. The function call links can be carried out to quickly gain a better understanding of how the source code sections relate to one another and which sections are either isolated or truly not relevant to the investigation.

Performing data element mapping alone may not sufficiently identify all relevant sections of source code. Most enterprise-level applications are designed with layers of abstraction for code reuse and ease of development. This is true for object-oriented programming languages, functional languages, and other modern language types. Relying on a method such as data element mapping will result in the investigator missing the ancillary sections of code that perform data operations. For example, the source code may have a main section of code that explicitly operates on several critical structured data fields, but those records are updated in a secondary portion of the source code that had no search matches in data element mapping.

Function call mapping is performed by identifying all possible mechanisms by which sections of source code can call or reference other sections of the source code and then searching for those keywords and referencing the results. The process is akin to data element mapping in that a set of keywords is created, the source code is searched based on those keywords, and the results are categorized and analyzed.

The first step, identifying all mechanisms for calling or referencing other sections of code, is conducted by creating a list of potential keywords specific to the programming language(s) in the source code and pattern indexes for how to identifying the reference. The investigator creates a list of keywords and regular expressions that will return the command that performed the calling or referencing and the section of code that was called. A common example in the C language is a function. The syntax will remain relatively consistent with parentheses after the function name and possible parameters within the parentheses. There are, however, additional mechanisms for calling or referencing additional sections of code. The source code could be stored as a .h "header" file that is referenced with the #include command. In addition, other code could be compiled and called as an external program via the ShellExecute() command. Regular expressions to store the calling command and referenced section of call need to be generated for any of the identified mechanisms.

The second step is to run the regular expressions from the previous step against the entire set of source code. The output should store the following:

 ISBN: 978-1-61208-242-4

- Calling/referencing mechanism;
- Called/referenced section of source code;
- Source code file where the match occurred, and
- Line of code where the match occurred.

The output from the second step next needs to be analyzed to identify key relationships and remove anomalies. Similar to data mapping, the first step is to quantify which source code objects are referenced most frequently. Examine the results to ensure those results are not false positives due to an incorrect regular expressions or a keyword that is used in other ways in the source code. The same is done for source code files that had too few or no search matches.

The findings from these steps can be shown in various ways. The simplest is simply identifying any sections of source code that are related critical sections of code. The critical code can be source code already known to be critical, results from the data element mapping, or source code that received a large number of search matches in the function call mapping phase. In addition, the findings can be depicted in graphic form to document the process flow for the source code. For example, Figure 1 was generated for a fraud investigation to show which AS/400 libraries either called or were called by the main source code library. Performing this analysis allowed the investigators to generate a map of all possible entry points for the enterprise investment trading platform

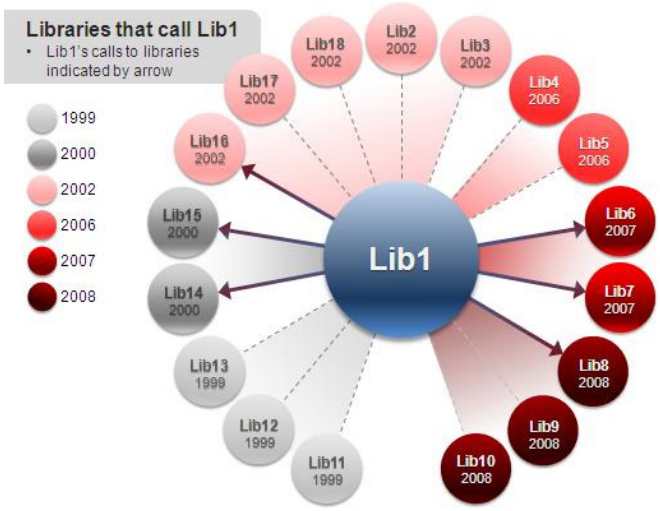


Figure 1. Function call mapping output.

Function call mapping is an iterative process that is typically run multiple times to refine the results. The critical sections of source code may contain additional mechanisms for calling or referencing other sections of source code, which requires the creational of additional regular expressions and searching the source code again. The amount of source code can affect the results, and in order to reduce the volume of search matches, investigators can limit their search population to key source code.

C. *Data Value Analysis*

The source code can be utilized to validate the data stored in the structured data. Most data in structured data repositories arrive via an external source and remained unchanged or were entered through automated logic. Data value analysis is performed to analyze for the latter. Under normal circumstances, structured data that are entered or updated through automated logic should be limited to only the types of values possible in the source code.

Acceptable value analysis is the process by which the constraints from the source code for particular fields are documented and then compared to the values in the structured data. In a fraud investigation, detecting anomalies is a key component for identifying areas in which fraud may have occurred. This analysis is typically performed only against key fields that are altered in the source code. Analyzing non-key fields is typically too time consuming. The process is performed by identifying how the data is altered in the source using the results from the data element mapping result set and then constructing a list of all unique values or patterns from the structured data set. A comparison will either show conformity or some anomaly. Anomalies may be caused by manual updates or data that were never altered by the source code. Both of these conditions require further analysis, as either case can be the result of fraud.

Unaltered field analysis is performed by analyzing all fields from the structured data that did not appear in the data element mapping process. Some fields may legitimately never be altered by the source code; however, if these fields appear in key tables, there is a possibility that that field is being manually updated. This analysis is performed by comparing the full field list for key tables against the data element mapping results for those tables. Any fields that do not appear in the data element mapping result set should be analyzed further.

Date value analysis, the third form of data value analysis, is a multi-step analysis based on either acceptable value analysis or external event analysis aimed at creating a chronology of the structured data and when the collected source code may have been operational. The first step is to identify the date range of records in key tables that have "unacceptable" values if there are any date values stored in the table or a linked, related table. These types of date fields are commonly stored in the table or in an audit table that stores a history of transactions. Since source code versions change over time, having a baseline date range is critical for knowing when the process for storing data in a particular table began. Next, identify any information about the source

 ISBN: 978-1-61208-242-4

code version history from embedded comments or related documentation. The investigator should analyze the data for acceptable values from the structured data and find the latest unacceptable value. That result is then compared to the source code change history documentation to identify any possible discrepancies.

External events can be analyzed in conjunction with the date value analysis by examining for higher volume of certain transactions or new transaction patterns. Most financial companies have fixed transaction patterns based on holidays, normal trading hours, and regulatory and legal events. The investigator should create a chronological list of important regulatory and legal events and normal trading dates. The normal trading dates should be compared to transaction volume from the structured data. In addition, that analysis should be compared to the acceptable value analysis to determine if the source code was modified in accordance with the regulatory and legal events. In many fraud cases, the source code and underlying data change dramatically when there is a regulatory or legal threat.

VI. CONCLUSION

Financial fraud investigations are becoming increasingly complex and require practical, agile approaches to reduce the volume to a manageable size. The volume of data continues to increase, as does the volume of underlying source code logic and the variance and interrelationships of the source code. As the volume and complexity increase, so too does the importance of identifying techniques for reducing the data to manageable sizes and identifying fraudulent activity as quickly as possible. Practical limitations make common theoretical approaches, such as semantic analysis, unfeasible and require techniques that do not depend on *a priori* knowledge of the source code and data.

Current research the related fields of data mining, source code reverse engineering, and fraud investigations provide useful tools and techniques that are appropriate for certain applications. Static analysis and UML-based approaches, when employed properly, can yield great results and provide useful insights in an automated fashion. Likewise, statistical data mining techniques allow an analyst to survey large volumes of data and understand the meaning and interrelationships of the data. The difficulty is that many of those techniques do not always have practical value when so many data variables are unknown, documentation is not available, and time constraints exist. Practical techniques where the meaning of the data and source code meet offer an alternative in those cases.

This paper presented the main practical techniques for ensuring that all data were properly collected and three critical methods for reducing the complexity of financial fraud investigations by identifying similarities and differences between the source code and structured data. Proper data collection is the vital first step for ensuring that all information is available for further analysis. By properly collecting and validating the data, one can be confident that the full investigation can commence. Data element mapping is a semi-automated method for reducing the volume of source code that needs to be analyzed when looking for a relationship between the source code and structured data. Function call mapping is a rapid method for identifying relationships between source code files and entry points in the application. Function call mapping is made more powerful by limiting the results to only related source code that have at least one section of code deemed critical to the investigation. Finally, data value analysis is a series of techniques to validate the contents of the structured data and identify potential anomalies.

The nature of financial fraud and the technologies used to conduct them continue to change, and as such, so too will fraud investigations change. Technological advancements and changing business practices, such as cloud computing and offshore data processing, introduce new complexity that will require advanced techniques for identifying critical information. So long as investigators remember to focus on identifying key data relationships and identify anomalies, advanced data analysis and visualization tools and techniques will allow investigators to be able to distill large volumes of information into their critical components and unravel the fraud.

REFERENCES

[1] U.S. Securities and Exchange Commission Office of Investigations, "Investigation of failure of the SEC to uncover Bernard Madoff's Ponzi scheme," August 31, 2009. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68–73.

[2] G. Canfora and M. DiPenta, "New Frontiers of Reverse Engineering," Future of Software Engineering, Future of Software Engineering (FOSE), 2007, pp. 326-341.

[3] A. Yazdanshenas and L. Moonen, "Crossing the Boundaries while Analyzing Heterogeneous Component-Based Software Systems," 27th IEEE International Conference on Software Maintenance, 2011.

[4] H. Mueller, J. Jahnke, D. Smith, M. Storey , S. Tilley, and K. Wong, "Reverse Engineering: A Roadmap," ICSE – Future of SE Track, 2000, pp. 47-60.

[5] L. Moonen, "Generating Robust Parsers using Island Grammars," Proceedings of the Working Conference on Reverse Engineering, 2001, pp. 13-22.

[6] W. Premerlani, M. Blaha, "An Approach for Reverse Engineering of Relational Databases," Communications of the ACM, Volume 37, Issue 5, May 1994, pp. 42-50.

[7] N. Mian, T. Hussain, "Database Reverse Engineering Tools," Proceedings of the 7th WSEAS International Conference on Software Engineering, February 2008, pp. 2006-2011.

[8] S. Rugaber and K. Stirewalt, "Model-Driven Reverse Engineering," IEEE Software Volume 21, 2004, pp. 45-53.

[9] F. Barbier, S. Eveillard, K. Youbi, O. Guitton, A. Perrier, E. Cariou, "Model Driven Reverse Engineering of COBOL," *Information System Transformations: Architecture Driven*

 ISBN: 978-1-61208-242-4

Modernization Case Studies, Mogran Kaufmann, 2010, pp. 283-299.

[10] ISO/IEC 19506:2012, "Information Technology – Object Management Group Architecture-Driven Modernization (ADM) – Knowledge Discovery Meta-Model (KDM), 2012.

[11] R. Kollmann, P. Selonen, E. Stroulia, "A Study on the Current State of the Art in Tool-Supported UML-based Static Reverse Engineering," Proceedings of the Ninth Working Conference on Reverse Engineering, 2002, pp. 22-32.

[12] I. Jacobson, "Ivar Jacobson on UML, MDA, and the Future of Methodologies," UML Forum FAQ,

[13] A. Sharma, P. Panigrahi, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," IJCA Journal, Volume 1, 2012, pp. 37-47.

[14] K. Fanning, K. Cogger, R. Srivastava, "Detection of Management Fraud: A Neureal Network Approach," International Journal of Intelligent Systems in Accounting, Finance, and Management, Volume 4, June 1995, pp. 113-126.

[15] S. Wang, "A Comprehensive Survey of Data Mining-Based Accounting – Fraud Detection Research," International Conference on Intelligent Computation Technology and Automation, Vollume 1, 2010, pp. 50-53.

[16] E. Kirkos, C. Spathis, Y. Manolopoulos, "Data Mining Techniques for the Detection of Fraudulent Financial Statements," Expert Systems with Applications, Volume 32, 2007, pp. 995-1003.

[17] J. Sremack, "The collection of large-scale structured data systems," Digital Evidence Magazine, January/February 2012.

[18] L. Hollaar, "Requesting and examining computer source code," in Expert Evidence Report, Vol. 4, No. 9, May 10, 2004, pp. 238-241.

[19] OWASP Foundation, "OWASP code review guide," Version 1.1, 2008, pp 49-50.

 ISBN: 978-1-61208-242-4

Integrity, or Lack Thereof, of the Electronic Record Systems of the Courts of the State of Israel

Joseph Zernik
Human Rights Alert (NGO)
Jerusalem
e-mail: 123456xtz@gmail.com

Abstract— **The Human Rights Alert (NGO) submission for the Universal Periodic Review of Human Rights in the State of Israel, filed in May 2012, is probably a first – being narrowly focused on analysis of integrity, or lack thereof, of the electronic record systems of the courts of the State of Israel, and being primarily based on data mining of this unique target area. Supreme Court: On or about March 2002, integrity of the electronic records was seriously compromised. Numerous fraudulent decision records were discovered. District Courts: The publicly accessible records were found invalid, primarily for failure to display visible, reliable digital signatures of judges and authentication records by clerks. Detainees Courts: The insecure, unsigned decisions of the detainees courts, often created long time after the dates of the hearings, could not possibly be considered valid legal records. The detainees ID numbers show suspicious discontinuities and failure to correlate with time of issuance, which should raise concerns regarding establishment of "black hole" prisons and "field courts". This study is a call for action by computing experts in general, and data mining experts in particular, in the safeguard of Human Rights and integrity of governments in the Digital Era.**

Keywords-e-courts; e-government; information systems; validation; electronic signatures; certification; authentication, State of Israel.

I. INTRODUCTION

The courts worldwide, including Israel, have been implementing in recent years electronic information systems for efficient management of court cases and public access to court records. Reports of the United Nations on Strengthening Judicial Integrity encourage this transition. Indeed, there is no doubt that such systems could improve the management of valid court records and transparency of the judicial processes. [1,2]

Court procedures, and in particular, the maintenance of valid court records, have evolved over centuries (in the case of the State of Israel – thousands of years, since the law of the State of Israel is in part based on Jewish Law) and are at the core of Fair Hearings.

The transition to electronic case management and public access systems, in any court, amounts to a sea change in these procedures. However, previous studies have shown that the transition to electronic record systems in the courts and prisons is not risk-free. [3-6]

This paper summarizes the results of a study of the electronic records systems of the courts of the State of Israel, which has been recently submitted for the January 2013 Universal Periodic Review of Human Rights in the State of Israel by the Human Rights Council of the United Nations (see Online Appendix 1). The information systems of the Supreme Court, the district courts, and the detainees' courts were examined, as well as records, pertaining to the implementation and enforcement of the *Electronic Signature Act* (2001).

The study presents the application of data mining techniques to a unique target area, and the ability of data mining techniques to analyze the integrity and validity, or lack thereof, of the systems, even under conditions, where the courts deny access to critical records, in apparent violation of the law.

II. THE LEGAL FOUNDATION FOR THE ADMINISTRATION OF THE OFFICE OF THE CLERK

While the primary responsibility of the judicial arm is in adjudication, the primary responsibility of the ministerial arm (clerks) is in the maintenance of honest court records, service and notice of judicial records, guaranteeing public access, and certification of judicial records. The authority, duties and responsibilities of the clerks and or registrars in the State of Israel were defined in a series of laws and respective regulations. The regulations, which were promulgated in 2003-5, during the period of implementation of the current electronic record systems of the courts are of particular interest. The laws and regulations are at times inconsistent in their basic terms and leave considerable amount of ambiguity in defining the procedures of the Office of the Clerk and the duties and responsibilities of the Chief Clerk and/or Registrar. On such legal background, it is also clear, that in developing and implementing the electronic record systems of the courts, the first step, e.g., defining the specifications of the electronic record systems, was particularly sensitive. Either the authorities, duties and responsibilities of the Chief Clerk and/or Registrar, and the respective procedures were to be unequivocally and unambiguously defined, or else an invalid electronic records system would be developed and implemented.

III. METHODS

The study was narrowly focused on analysis of integrity of the electronic record systems in the national courts (Supreme Court, district courts, detainees' courts). The study was not based on legal analysis of these records, or challenges to the rationale of the adjudication, except for the laws pertaining to the maintenance of court records. Instead, irregularities in date, signature, certification, and registration

 ISBN: 978-1-61208-242-4

procedures were examined through data mining methods, executed on the online public records of the courts.

Initially, integrity of the basic components of the systems was examined: indices of all cases, calendars, dockets (lists of records in a given file), indices of decisions, and compliance of these components with the *Regulations of the Courts,* pertaining to the maintenance of court records. and consistency of data among these components (e.g., the date a record was filed, as listed in the docket, and as indicated in the body of the record itself, see Online Appendix for links to data).

Subsequently, a cursory survey was conducted of the pattern of judges' signatures and clerk's certification of records over the past two decades. The significance of events around 2001-2003 was identified.

Accordingly, a more detailed survey was conducted of records of that period, including data mining relative to changes in distribution of specific word combinations, related to certification over time (e.g., "Chief Clerk", "Registrar", "Shmaryahu Cohen" (the late Chief Clerk of the Supreme Court), "Boaz Okon" (former Registrar of the Supreme Court), True Copy).

Subsequently, court records that were identified as outliers in such distributions (e.g. Decision records bearing the name of the late Chief Clerk Shmaryahu Cohen, issued later than the date of his death) were individually examined. Such data mining procedures enabled the discovery of hundreds of fraudulent decision records.

Once the death of the late Chief Clerk of the Supreme Court on March 7, 2002, was identified as a key event in this context, Google searches were conducted to further elucidate the event. It turned out that he reportedly died of "sudden cardiac arrest", after toasting a retiring staff member in an office party. Additionally, Google searches discovered a complaint, filed with the Israel Police by a family member/friend two weeks after the event, alleging murder. However, the complaint failed to present any reasonable motive for such murder. Regardless, web pages were discovered with various conspiracy theories in this regard.

Based on the findings from such data mining efforts, requests were filed on the Ministry of Justice and the Administration of Courts, pursuant to the *Freedom of Information Act*, for records that would provide the legal foundation for the profound changes in certification patterns between 2001-3, the appointment records of the current chief clerks of the courts, the appointment records of the Registrars of Certifying Authorities, pursuant to the *Electronic Signature Act* (2001), secondary legislation that might have authorized the changes, etc.

Additionally, outside sources were reviewed for information regarding the history of the development and implementation of the electronic records systems of the courts: media reports, and in particular the 2010 *State Ombudsman's Report* 60b.

The analysis was also based on consultations with Israeli law and computing/cryptology experts.

IV. THE SUPREME COURT

The March 7, 2002 untimely death of Chief Clerk of the Supreme Court Shmaryahu Cohen is tightly correlated with precipitous corruption of the electronic records of the Supreme Court. Today, identity of the Supreme Court's servers is not verified, and all Supreme Court decisions are published unsigned by judges, uncertified by the clerks, and subject to "editing and phrasing changes" (Figure 1). In the transition period (2001-2003), numerous Supreme Court decisions were falsified (Figure 2).

Today, the Supreme Court refuses to comply with the law regarding service of its decisions by the Clerk and denies public access to the authentication records – the certificates of delivery, even to a party in his/her own case.

False and deliberately misleading certifications of Supreme Court decisions, recently issued by the office of the Chief Clerk were also discovered.

Effectively, the Supreme Court established a 'triple-book' record system, where the public and parties to litigation are not permitted to distinguish between valid and simulated, i.e., fraudulent decisions (see Online Appendix for further details and links to data).

V. DISTRICT COURTS

The evidence shows that implementation of *Net Ha-Mishpat*, the electronic record system of the district courts, undermined the integrity of the records of the courts, and in particular, the accountability of the Chief Clerks relative to the integrity of the records.

The 2010 *State Ombudsman's Report* 60b reviewed the development and implementation of *Net Ha-Mishpat*. [7] The report describes a system that was developed with no written specification and with no core supervision by State employees, the issuance of contracts to outside corporations with no bidding, and acceptance of the system with no independent testing of its performance by State employees. Most alarming, the Ombudsman's Report indicated that unknown number of individuals had been issued double Smart ID cards. The Ombudsman pointed out that the development and implementation of the system was conducted in violation of State law. However, the report failed to evaluate the validity of the system as a whole.

The records of the district courts, which are publicly accessible in Net Ha-Mishpat, cannot possibly be deemed valid legal records (Figure 3).

News media reports revealed material conflicts of interests by individuals, who were in key positions, relative to the development and implementation of *Net Ha-Mishpat*.

VI. DETAINEES' COURTS

Analysis of the detainees' ID numbers, as they appear in the online publicly accessible records, showed discontinuity in the ID numbers, and lack of correlation between the ID numbers and date of issuance (Figure 4).

The refusal of the Ministry of Justice to disclose the number and locations of such courts in response to a Freedom of Information request, combined with the invalid

 ISBN: 978-1-61208-242-4

Detainee Numbers should raise concerns that 'black hole' prisons with makeshift 'field courts' have been established.

Review of news media revealed numerous reports of abuse of Due Process in the detainees' courts. In 2010 Haaretz daily reported the conduct of a simulated hearing and the issuance of simulated court order in the case of a detainee. [8] Haaretz quoted the spokesperson of the Ministry of Justice, referring to the report as "a tempest in a teapot", and claiming that the case was only a secretarial error in data entry.

VII. COMPLIANCE WITH STATE OF ISRAEL LAWS

The evidence shows that development and implementation of the new electronic record systems of the courts should be deemed in violation of the law of the State of Israel.

A. The Israeli "Constitutional Revolution"

The State of Israel has not established a constitution to this date. In the early 1990s, the Knesset (legislature) enacted two "Basic Laws", in effort to establish fundamental Human Rights by law. Moreover, under the tenure of Presiding Justice Aharon Barak (1995-2006), and to a lesser degree under the tenure of Presiding Justice Dorit Beinisch (2006-2012) the Supreme Court purportedly spear headed a "Constitutional Revolution" (see also Online Appendix). [9] Various "Constitutional Rights" were purportedly construed by the Supreme Court, e.g., in *Israeli Civil Rights Association v Minister of Justice* (5917/97).

On the other hand, the current study documents precipitous corruption of the courts of the State of Israel, in particular – the Supreme Court, during the very same years that the Supreme Court's rhetoric regarding "Constitutional Revolution" and "Constitutional Rights" reached its zenith.

B. Basic Law – Human Dignity and Liberty (1992)

The *Basic Law – Human Dignity and Liberty* (1992), together with academic papers by former Presiding Justice Aharon Barak on the subject, are often credited with launching the "Constitutional Revolution". However, some legal experts opined that the Basic Law is vague and ambiguous. To the degree that Basic Law guarantees Human Rights, such as Due Process/Fair Hearings, Access to Justice and national tribunals for protections of rights, the current study documents that the Supreme Court disregards the Basic Law.

C. Laws and regulations pertaining to the administration of the courts

The Chief Clerk of the Supreme Court did not comply with the *Regulations of the Courts – Office of the Clerk* (2004), and provided fraudulent certification of Supreme Court decisions (Figure 5).

The Supreme Court's refusal to duly serve and authenticate its own decisions, was also out of compliance with the Supreme Court's decision in *Israeli Bar Association v Minister of Religious Affairs et al* (6112/02).

Additionally, the Supreme Court's refusal to permit a party to inspect court records in his own case, was also out of compliance with the Supreme Court's decision in *Israeli Bar Association v Minister of Justice* (5917/97).

The 2010 *State Ombudsman's Report* 60b also documented violation of the *Regulations of the Courts – Office of the Clerk* (2004), relative to the removal of the servers of the electronic records from the offices of the Clerks of the Courts onto corporate grounds.

The fraudulent Apostille certification procedure, published online by the Administration of Courts documents violation of the *Regulations of the Courts – Office of the Clerk* (2004), since the procedure purports to permit notaries to certify court records, which the regulations authorize only the Chief Clerks to certify. Both the Administration of Courts and the Ministry of Justice refused to respond on the Freedom of Information requests, pertaining to legal foundation of the Apostille certification procedure and the identity of those, who authorized its publication.

The refusal of the Administration of Courts to produce the appointment records of the Chief Clerk of the Supreme Court and the district courts, should raise concerns regarding the lawful nature of their appointments.

Together, the use of servers, whose identity is not certified, the publication of court decisions, which are neither signed, nor certified by the clerk, and dubious accountability of those, who serve today as chief clerks of the courts, conditions were set, where integrity of the electronic records of the State of Israel should be deemed dubious at best.

With it, the study identified an abundance of falsified court records, and simulated, illegal public records (e.g. the Apostille certification procedure), that were published online by the courts and the Ministry of Justice.

D. Electronic Signature Act (2001)

The Act and the respective regulations were signed and became effective in 2001. In pertinent parts, the Act says:

> Chapter 2. Validity of a Secure Electronic Signature
> …
> 2. (a) For any law, requiring a signature on a document – such requirement may be fulfilled, in respect of any electronic message, by use of an electronic signature, provided that it is a certified electronic signature…
> 3. An electronic message, signed with a secure electronic signature is admissible in any legal procedure…
> 4. A certified electronic signature is presumed to be a secure electronic signature.

Pursuant to the Act, a Registrar of Certification Authorities (qualified as a district judge) was to be appointed in the Ministry of Justice. Discontinuities in certification authorities of the Supreme Court in late 2001-2, which ended with no certification at all, followed closely, and were possibly related to the signing of the Act and regulations. Several individuals purportedly served in the

 ISBN: 978-1-61208-242-4

position of Registrar over the past decade, Guidelines and Standards were published, and enforcement was conducted. In 2009, Director of newly minted "Law, Technology, and Information Authority" was appointed, as part of reorganization in the Ministry of Justice. It appears that since then the position of Registrar ceased to exist (see additional details in the Online Appendix).

Requests, *pursuant to the Freedom of Information Act* (1988), pertaining to the implementation of the *Electronic Signature Act* and appointment records of the Registrars of Certifying Authorities of the past decade, were not answered by the Ministry of Justice and the Administration of Court, or answered in a manner that should be deemed false and deliberately misleading.

Combined, the evidence shows that the Ministry of Justice has deliberately undermined the implementation of *Electronic Signature Act* (2001) over the past decade.

E. *Freedom of Information Act (1988)*

Both the Administration of Courts and the Ministry of Justice refused to answer, or provided invalid, or false and deliberately misleading responses on Freedom of Information requests, pertaining to integrity of the electronic records of the courts, e.g., legal records, which would provide the foundation for the changes in certification practices in the Supreme Court, or the Apostille certification procedures, appointment records of the chief clerks of the courts, appointment records of Registrar of Certifying Authorities, the identities of any Certifying Authorities that may have been assigned to the courts, names of individuals, who hold the ultimate administrative authority for the servers of the courts, the names and locations of the detainees' courts, etc (see Online Appendix for complete log of Freedom of Information requests and responses).

VIII. COMPLIANCE WITH RELEVANT TREATIES AND CONVENTIONS

The *Hague Convention* (1961), to which the State of Israel is a party, established an Apostille certification procedure, in order to validate legal public records, which are taken from the courts of one nation to another. The Apostille certification procedure, which was published online by the Administration of Courts, unsigned, undated, and with no reference to any legal foundation, is opined as a deliberate effort to undermine the integrity of Apostilles, originating in the courts of the State of Israel (Figure 5).

IX. CONCLUSIONS AND RECOMMENDATIONS

The results of the current study show that senior members of the judiciary and the legal profession exploited the transition to new electronic record systems in the courts of the State of Israel over the past decade to undermine the integrity of the justice system. It appears that updates in the electronic records systems and the passage of the *Electronic Signature Act* made it necessary to decide between the development of systems, based on valid, lawful specifications and lawful digital signatures, or systems based on no specifications and no digital signatures at all.

The results show that effectively, decision was made around 2002 in favor of the latter option. The most obvious trait of the systems now in place, is that among thousands of electronic public legal records, which were examined as part of the current study, not a single digitally signed record was discovered.

Furthermore, the findings suggest that such decision required the neutralization of the main watchdogs, relative to integrity of legal records: the chief clerks of the Supreme Court and the district courts, and the Registrar of Certifying Authorities.

It was also necessary to devise ways to circumvent the valid certification procedures, still in existence in paper form, as documented in the fraudulent certifications by the office of the Chief Clerk of the Supreme Court (see Online Appendix for figures an data), and the fraudulent Apostille certification procedure (Figure 5).

Finally, although the online publication of court records could have increased public access and transparency of the courts, ways were devised, whereby the online public access system would not permit the public to distinguish between valid and void court records. Separate data bases were concealed in the case management system of the courts, where public access is denied. Therefore, the perception of public access was created, while in fact public access and transparency of the courts were undermined. [10]

The outcome, best documented in the Supreme Court and the Detainees' Courts, was the enabling of the publication of simulated court decisions and conduct of simulate court proceedings.

The resulting conditions are likely to lead to deterioration in the Human Rights of the People of the State of Israel, albeit, some years may pass before the full impact is manifested. Furthermore, the failure to uphold the *Electronic Signature Act* has ramifications far beyond the justice system. It is likely to place Israeli financial markets at high-risk of instability.

Therefore, the implementation of invalid electronic record systems in the courts holds serious implications relative to Human Rights and banking regulation in the State of Israel. (see Online Appendix for further details).

The findings should also require reassessment of any faith and credit, which may be given to legal public records originating in the courts of the State of Israel by other nations, including, but not limited to those, who are parties to the *Hague Convention* (1961).

The findings hold serious implications relative to Human Rights and banking regulation in the State of Israel.

The Human Rights Alert submission recommends:

1. The electronic records systems of the courts should be examined and repaired by Israeli computing and legal experts, under accountability to the legislature.

2. A Truth and Reconciliation Commission should be established to examine the conduct of members of the judiciary and the legal profession, who were involved in undermining the integrity of the electronic record systems;

 ISBN: 978-1-61208-242-4

3. No court of any nation should be permitted to develop and implement its own electronic record systems, since such systems effectively amount to establishment of new regulation of the courts. Typically, the authority to establish such regulations is reserved for one of the other two branches of government.

The results of the current study are not unique to the State of Israel. The Human Rights Alert submission for the 2010 Universal Periodic Review of Human Rights in the United States was in part based on analysis of the lack of integrity in the electronic record systems of the California courts and prisons. The submission was reviewed by the United Nations professional staff and incorporated into the official report with a note referring to "corruption of the courts and the legal profession in California". [11] An accompanying paper describes the fraud inherent in the electronic record systems of the courts of the United States, which were implemented a decade earlier than the systems described in this study. Preliminary inspection suggests that similar faults also exist in the electronic record systems, which have been recently implemented in other "Western Democracies".

Finally, this study is a call for action by computing experts in general, and data mining experts in particular, in the safeguard of Human Rights and integrity of governments in the Digital Era.

ACKNOWLEDGMENT

The author thanks Israeli computing/cryptology and legal experts for their assistance.

ONLINE APPENDIX

[1] The online appendix includes further details, links to the original data, and enlarged and additional figures:
Human Right Alert (NOG), 2013 State of Israel UPR Appendix to Submission: "Integrity, or lack thereof, in the electronic record systems of the courts of the State of Israel" http://www.scribd.com/doc/82927700/

REFERENCES

[1] United Nations Drug Control and Crime Prevention Center, "Report of the First Vienna Convention Strengthening Judicial Integrity", CICP-6, 2000

[2] United Nations Drug Control and Crime Prevention Center, "Strengthening Judicial Integrity Against Corruption", CICP-10, 2001

[3] J. Zernik, "Data Mining as a Civic Duty – Online Public Prisoners' Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining 1: 84-96, 2010

[4] J. Zernik, "Data Mining of Online Judicial Records of the Networked US Federal Courts", International Journal on Social Media: Monitoring, Measurement, Mining, 1:69-83, 2010

[5] J. Zernik, "Fraud and corruption in the US courts is tightly linked to failing banking regulation and the financial crisis", Proceedings of the 16th World Criminology Congress, Kobe, Japan, 345, 2011

[6] J. Zernik, "Invalid case management and public access systems, implemented in the Israeli courts, undermine Human Rights", Proceedings of the 16th World Criminology Congress, Kobe, Japan, 348, 2011

[7] State of Israel Ombudman, "Report 60b (2010) - Ministry of Justice Computerization pp 693-747, 2010

[8] D. Weiler, "Court Decision Appeared in Court File Prior to Proceedings", Haaretz, February 8, 2011

[9] A. Barak, "A Constitutional Revolution – Israel's Basic Laws", Constitutional Forum 4:83-85, 1992

[10] N. Sharvit, Chief Justice Dorit Beinish: The new case management system (Net Ha-Mishpat) would require restricting public access to court records, Globes, October 8, 2009

[11] Human Rights Council, Working Group on the Universal Periodic Review, Ninth session, "National report submitted in accordance with paragraph 15 (a) of the annex to Human Rights Council resolution 5/1, United States of America" (1–12 November, 2010

Figure 1. Changes in the Supreme Court's Chief Clerk's Certification of Electronic Decision Records of the Supreme Court of the State of Israel Between 2001-2003

(a) Until early 2002, all electronic decisions of the Supreme Court carried certification by the late Chief Clerk Shmaryahu Cohen. (b) Since 2003, none of the electronic decision records carries any certification, or any reference to the Office of the Clerk. Instead they carry a disclaimer "subject to editing and phrasing changes", and reference to an "Information Center", which has no foundation in the law. The Administration of Courts refuses to disclose, in response to Freedom of Information request, the legal foundation for such profound change in the records of the Supreme Court in 2001-2003.

 ISBN: 978-1-61208-242-4


ניתנה היום, כ"ו שבט, תשס"ז (14/02/2007).

בעז אוקון

רשם

נוסח זה כפוף לשינויי עריכה וניסוח.
שמריהו כהן - מזכיר ראשי
בבית המשפט העליון פועל מרכז מידע, טל' 02-6750444
בית המשפט פתוח להערות והצעות: pniot@supreme.court.gov.il
לבתי המשפט אתר באינטרנט: www.court.gov.il

Figure 2. Fraud in Electronic Decision Record of the Supreme Court of the State of Israel.

The Decision, in *Judith Franco Sidi et al v Authority pursuant to the Persons Disabled by Nazi Persecutions Act* (1582/02) in the Supreme Court in part says:

> Issued this date, February 14, 2007
> Boaz Okon
> Registrar
>
> _____
> This version is subject to editing and phrasing changes.
> Shmaryahu Cohen – Chief Clerk
> In the Supreme Court an information center is operated, Tel: 02-6750444
> The Court is open to comments and suggestions: pniot@supreme.court.gov.il
> The courts' web site: www.court.gov.il

By February 2007, Boaz Okon was no longer Registrar of the Supreme Court, and Shmaryahu Cohen was dead for about five years. Numerous other records of the same nature were discovered.

 ISBN: 978-1-61208-242-4

Figure 3. Invalid Electronic "Post-it" Decision Record of the Jerusalem District Court

The record belongs to the file of *State of Israel v Awisat et al* (9739-01-11) in the Jerusalem District Court. The background document, is a motion, titled: "Request by Joint Stipulation for Continuation of Hearing Date." The Motion record is graphically signed by the attorney who filed the motion in the lower left corner. The small framed image, superimposed on the Motion records in the upper right corner, is a "Post-it Decision". The yellow heading states:

January 25, 2011
Judge Amnon Cohen, Decision

The framed text below the yellow heading states:

The Request is denied. Moreover, it was filed out of compliance with the Guidelines of the Presiding Judge.

The decision is neither visibly signed by Judge Cohen, nor certified by the Clerk of the Court, and it fails to bear the seal of the Court. No visible electronic signatures, pursuant to the *Electronic Signature Act* (2001) were implemented in the electronic records systems of the courts of the State of Israel. The Administration of Courts refuses to produced the appointment records of the Chief Clerks of the district courts, or to disclose, who holds the ultimate administrative authority over the electronic records of the district courts.

 ISBN: 978-1-61208-242-4

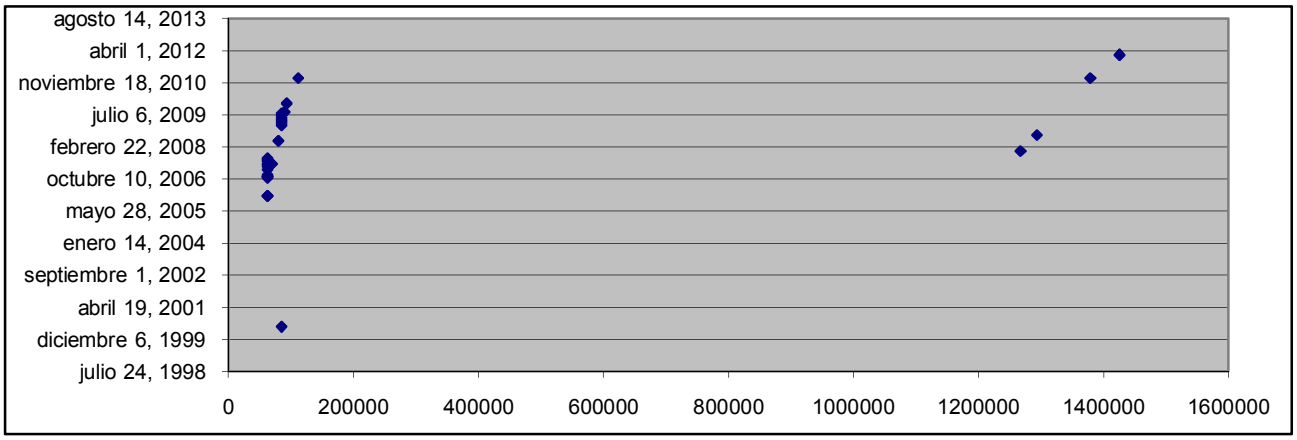


Figure 4. Lack of Integrity in Detainees' ID Numbers, in Records of the Detainees Courts of the State of Israel

The lack of correlation between dates of issuance of the decisions, and Detainee Numbers, and the apparent discontinuity in Detainees' ID numbers, should be deemed a fundamental failure of integrity of the Detainees Courts electronic record system. (See the raw data at Table 3 in the online appendix). Only a selection of the Detainees Courts records is published online, as insecure Word files, most of which were created a long time after the fact (at times – years). The Ministry of Justice refuses to disclose, how many Detainees Courts are operating in the State of Israel, their names and locations, and the names of the Chief Clerks of the detainees courts, if any exist. Combined, the findings should raise concern that "black hole" prisons and makeshift "field courts" have been established in the State of Israel.

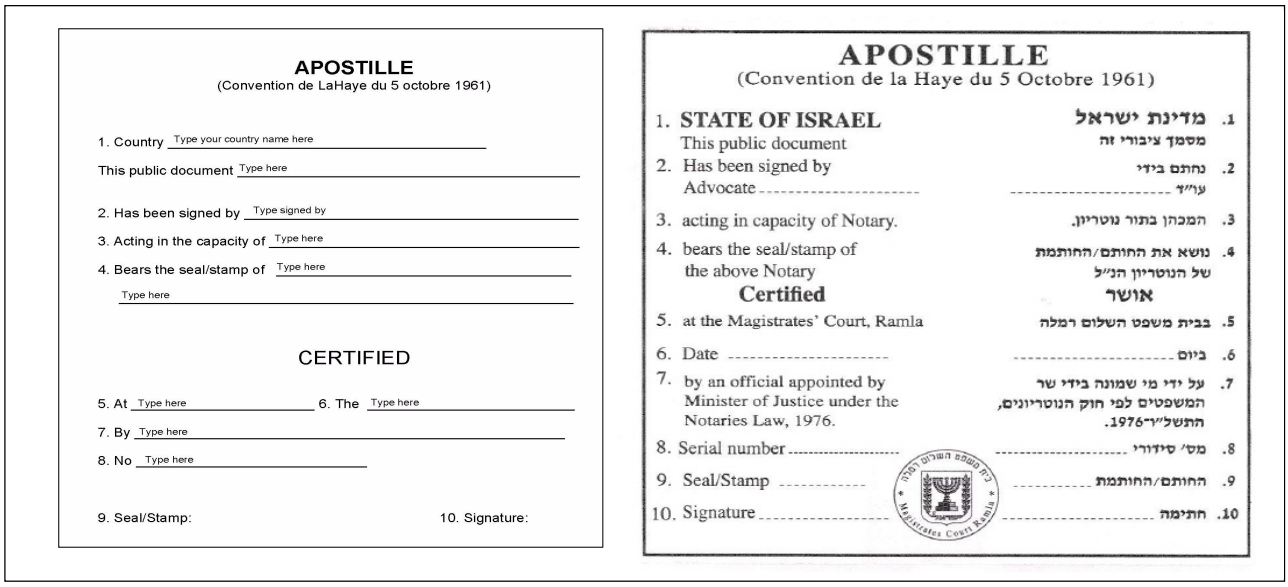


Figure 5. Fraud in Apostille Certification Procedure, Pursuant to the *Hague Apostille Convention* (1961), Published Online by the "Judicial Authority"

Left: True apostille form, as authorized by the *Hague Apostille Convention* (1961); Right: A sample apostille form, published on the web site of the "Judicial Authority" of the State of Israel, falsely represented as the true apostille form, as authorized by the Convention. The form, published by the "Judicial Authority", purports that an "Advocate", acting as a Notary, is permitted to certify court decisions, which the *Regulations of the Courts – Office of the Clerk* (2004) authorize only the Chief Clerks to certify. Furthermore, the latter form permits a member of the staff of the office of the clerk, to sign the apostille form, as certification of the signature of the Notary, with the Seal of the Court, in a manner that appears as a valid certification by a chief clerk of the attached court decision. In fact, the arrangement, published online, specifically states that in executing the apostille, the office of the clerk certified ONLY the signature of the notary, but not the attached court record. The arrangement is opined as fraud on the People of the State of Israel, and also on the People and the courts of all other nations, who are parties to the Convention. It is part of a pattern of false certifications of records of the courts of the State of Israel. Both the Administration of Courts and the Ministry of Justice refuse to disclose, who authorized this procedure, and who and when authorized its online publication. The Chief Clerk of the Supreme Court refused to provide apostille certification of judicial records of the Supreme Court.

 ISBN: 978-1-61208-242-4

Network Visualization of Car Inspection Data using Graph Layout

Jaakko Talonen, Miki Sirola
Aalto University, Department of Information and Computer Science
Espoo, Finland
jaakko.talonen@aalto.fi, miki.sirola@aalto.fi

Mika Sulkava
MTT Agrifood Research Finland, Economic Research
Helsinki, Finland
mika.sulkava@mtt.fi

***Abstract*—In this paper, we introduce the network visualization based on the rejection reasons on car inspections. It is compared with the visualization based on principal component analysis. The largest private provider of vehicle inspections in northern Europe, A-Katsastus, published rejection statistics in Finland for the fourth time. The statistics is published in dozens of tables on the basis of the year of introduction into use, make or model. Our goal is to visualize all this information in one network. The car inspection data is aggregated with the produced visualization. However, the dependencies between the different rejection reasons and cars can be efficiently studied by exploring our network visualization.**

Keywords-Car inspection; Rejection reason network; Visualization; Gephi; ForceAtlas2

I. INTRODUCTION

In our research, we used a desktop application Gephi. It is an interactive visualization and exploration platform [1]. It is commonly used for the visualization of social networks [2], Facebook friends, Twitter, etc. For example, the discussions between the Members of the Parliament in Finland have been visualized [3]. In that visualization, those who took part to the same discussions are connected. Visualizations of Social Networks have been widely studied, e.g., in [4], a network of social relations is created using publicly articulated mutual "friendship" links. In this paper, we present results suggesting that network visualization is suitable also for car inspection data where different rejection reasons are presented as "bridges" between the cars.

Typically, in reliability rating reports, areas of inspection are excluded and the most common grounds for rejection to keep legibility simple enough on the different tables, e.g., TÜV reports [5]. Only rank, car make – model and fault percentage in different "introduced into use" tables are shown. It means that there are 40 different tables published per year, i.e., 320 tables since 2004. Our goal is to aggregate all this data to one visualization including the most common grounds for rejection. It helps the users to make their own conclusions by extracting some rejection reasons, such as tires, which are totally dependent on drivers when analyzing the data. Our network visualization is compared with a visualization generated by Principal Component Analysis (PCA) which was used on our earlier work [6].

II. DATA AND PREPROCESSING

A-Katsastus Group inspected approximately 925 000 passenger cars in Finland in 2011 [7]. This and previously published data is used in our visualization. Yearly data is divided into several tables by the age of the car. A similar publication has previously been produced on the basis of their statistics by the Swedish company Bilprovningen and the German vehicle inspection chain Dekra.

The top three rejection reasons (RR) are listed, if a certain car is inspected more than $N=100$ times and the same rejection reason is listed more than $L=10$ times. The proportion of these requirements is $p=0.1$. In theory $p \in]0,1]$, but in practice it can be assumed that p varies around value 0.1. In this paper "car" means its model, type and age. In Finland, new cars are inspected on third and fifth year and older cars yearly. Newer cars have fewer rejections. Therefore there is less information about new cars than old ones. Also an average rejection r [percentage] is listed.

In the original data rejections are divided into 13 different classes, such as tires, brakes, steering and control devices; see Table 1. For analysis, a 14th class is defined as "unclassified" reason, because mostly the new cars have no listed rejection reasons in the data.

TABLE I. REJECTION REASONS

Classified rejection reasons and the sum of most popular reasons for rejection [A-Katsastus]		
RR	***ID***	***# 1st RR (2011)***
chassis	1	9
front suspension	2	73
shock absorption	3	17
suspension	4	12
brakes	5	163
other equipment	6	0
steering and control devices	7	53
exhaust emissions	8	102
tires	9	0
parking brake	10	22
rear axle	11	18
airbags	12	0
identification number	13	0

 ISBN: 978-1-61208-242-4

Data is quantified for the analysis using the rejection percentage r and the rejection reasons. We define a car matrix as

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} & x_{1,m+1} \\ \vdots & \ddots & & \\ x_{n,1} & & x_{n,m} & x_{n,m+1} \end{bmatrix}, \quad (1)$$

where n is the number of different cars and m the number of classes. In quantification it is assumed that the k^{th} RR j has probability

$$x_{i,j} = (p + a(k)) \cdot r, \quad (2)$$

where vector a is defined in this research as:

- All three reasons are listed: a = [0.04 0.02 0],
- Two reasons are listed: a = [0.02 0],
- Only one reason is listed: a = 0.

So, if all three rejection reasons are listed for the car i, the row sum of the first 13 cells for this car is $(0.14+0.12+0.1)r = 0.36r$ with the assumption $p=0.1$ (based on the A-Katsastus publication requirements). Last cell $x(i,14) = 0.64r$, so the sum of each row X is r. If no reasons are listed for a car, then $x(i,14) = r$. In practice this means that we assumed that the most common (listed first) reason is $2r$ percentage units more probable than the second listed reason.

In the year 2008, RRs were classified in a different way than in the years 2009 - 2011. Therefore, it was excluded from the analysis and three matrices X(2009), X(2010) and X(2011) were visualized. As mentioned before, newer cars are inspected every second year, so in matrices X(2009) and X(2011) there are cars, which are missing from the matrix X(2010). These data matrices were combined row by row. In a new matrix Z one row represents one car. Older statistics have less effect on the model and it is defined as

$$Z(i,:) = \frac{\sum_{k=2009}^{2011} \lambda^{2011-k} L(i,2011-k) X_k(i,:)}{L(i,1) + L(i,2)\lambda + L(i,3)\lambda^2}, \quad (3)$$

where $\lambda=[0,1]$ is a forgetting factor. If $\lambda=\alpha$, it is expected that α% of last years car RRs are taken into account in the visualization. L is a zero-one vector defining if car data exists on matrix X(k) or not. Car i introduced in use, e.g., in 2007 was inspected in the year 2010, but not in the years 2009 and 2011, so then $L(i,:) = [0\ 1\ 0]$. Zero values in matrix Z mean that there is no connection between the car i and RR j.

III. METHODS

A. Principal Component Analysis

Principal Component Analysis (PCA) is a method for orthogonal linear transformation. The dimension of the data is reduced by transforming it to a new coordinate system such that the greatest variance lies on the first component [9]. The quantified matrices are combined as

$$C = \begin{bmatrix} X_{2009} & X_{2010} & X_{2011} \end{bmatrix}^T, \quad (4)$$

and matrix C is projected to subspace by placing the first N principal components in matrix

$$\Theta = (\theta_1 | \cdots | \theta_N). \quad (5)$$

B. Network Visualization

Our network layout is based on the ForceAtlas2 (FA2) method [8]. It is suitable for graphs with 10 to 10000 nodes. Cars and RRs are represented by colored balls in a graph. The attraction force F between two nodes $n(1)$ and $n(2)$ depends linearly on the distance $d(n(1),n(2))$.

FA2 is a continuous algorithm and the model is based on attraction and repulsion proportional to distance between nodes. Various layouts are achieved with different initial coordinates and parameter settings, see (2, 3). The main goal is to produce a readable spatialization and devise an energy model that could be easily understood by users. A clear visualization where nodes are separated and not overlapped (this feature can be forced in Gephi [1]) is reached by various parameter settings. In our model default values of scaling were increased and gravity decreased to obtain a sparser graph.

IV. EXPERIMENTS

A. Principal Component Analysis

We tested two different methods for car data visualization. Matrix C was projected to a 2-dimensional space using PCA. First and second components were visualized using Google Motion Chart where the RR statistics for 2009-2011 can be interactively explored. Cars with the same RR are situated in the same place in coordinates. However, this visualization is not very practical. Cars with only one listed RR are situated in corners, because m in (1) relatively small [6]. Readibility of the graph is not very good if only one plot is used, because RRs are shown only in a loadings plot, see Figure 1.

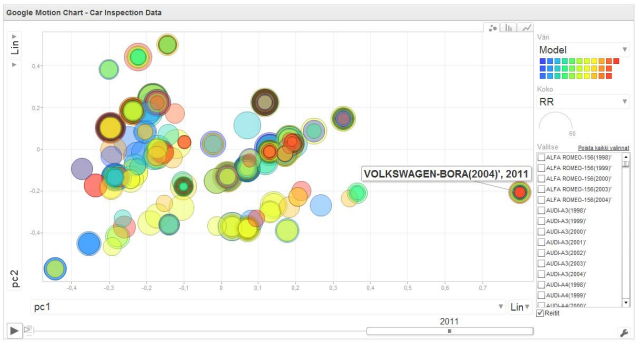


Figure 1. PCA results are shown in the motion chart. Volkswagen Bora (2004) in the last year's statistics (2011) is projected to the far right. The first component loading for exhaust problem is positive.

B. Network Visualization

All cars are connected to rejection results based on matrix Z, see (3). ForceAtlas2 algorithm is used to order initial car coordinates to stabile positions. Some of the

default parameters were changed: *Edge Weight Influence = 0.02, Scaling = 50, Gravity = 1.0 and Tolerance = 0.1*. Cars with same rejection reasons are situated in the same area in the graph.

Some cars have no listed rejection reasons. Therefore, the unclassified class *x(i,14)* was introduced in the previous section. Cars without RR information are located in the same area. The total weights of RRs are the same as the average rejection percentage *r*. Each car type and model is connected to the same cars with up to two years difference in age with small connection weights. With one year difference *w=0.2* and two years difference with *w=0.1*. By this procedure, the visualization is more informative, because a relatively large cluster of nodes with only one connection to unclassified RR is avoided. In practice, it means that in a graph we are assuming that if car *i* does not have listed RRs, it has with small probability the same RRs than older or newer cars with the same model and type, but are not listed in the published statistics.

In the past three years, there were 1060 cars which were inspected more than 100 times. It means that the size of matrix Z is *(n=1060, m=14)* in our visualizations. A produced network with *(λ = 0.8)* is visualized in Figure 2. All cars are connected to unclassified RR (ID=14) which is situated in the center of a graph. Infrequent RRs are placed on graph borders and RR(ID=i) which is dependent with another RR(ID=j) near each other.

The parking brake RR is mentioned in car inspection statistics in highlighted cars shown in Figure 3. Chassis, front suspension and brakes problems also occur rather probably in these cars, because these RRs are rather close each other in the graph.

Some of the RRs are totally driver dependent. Rather old Toyota Corollas have had small rejection percentage *r*. In addition, one of the reasons was bald tires. The connections of Toyota Corolla (2006) are shown in Figure 4. About *3.5%* of the cars were rejected in the car inspections. Based on matrix Z, connections show that in the past three years 3rd top-rated reason was tires. Also, additional connections show that older and newer cars have had the same kind of RRs.

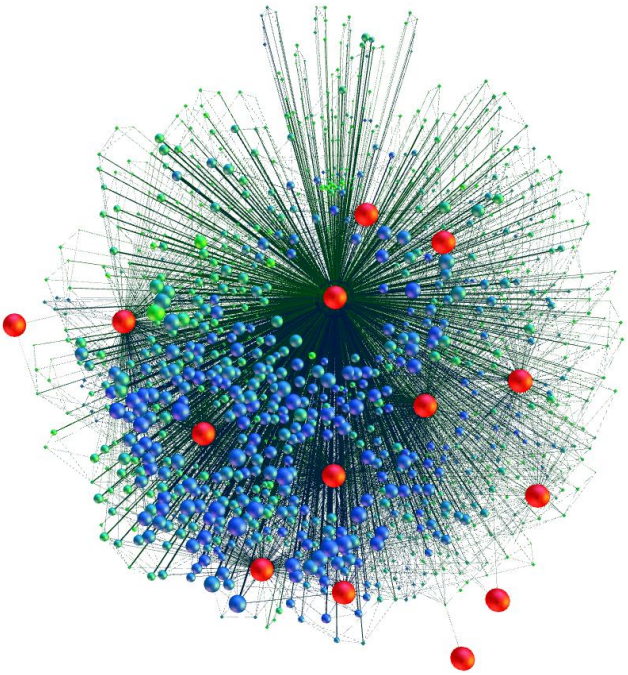


Figure 2. Cars and rejection reason classes (constant size red balls, m=14) are visualized. Old cars are represented by blue balls and newest (2008) by green balls, n=1060. Edges exist between RRs and cars. Also cars with age ±2 are connected. Sizes of the car balls and edges between cars and RRs are proportional with the rejection rates *r*.

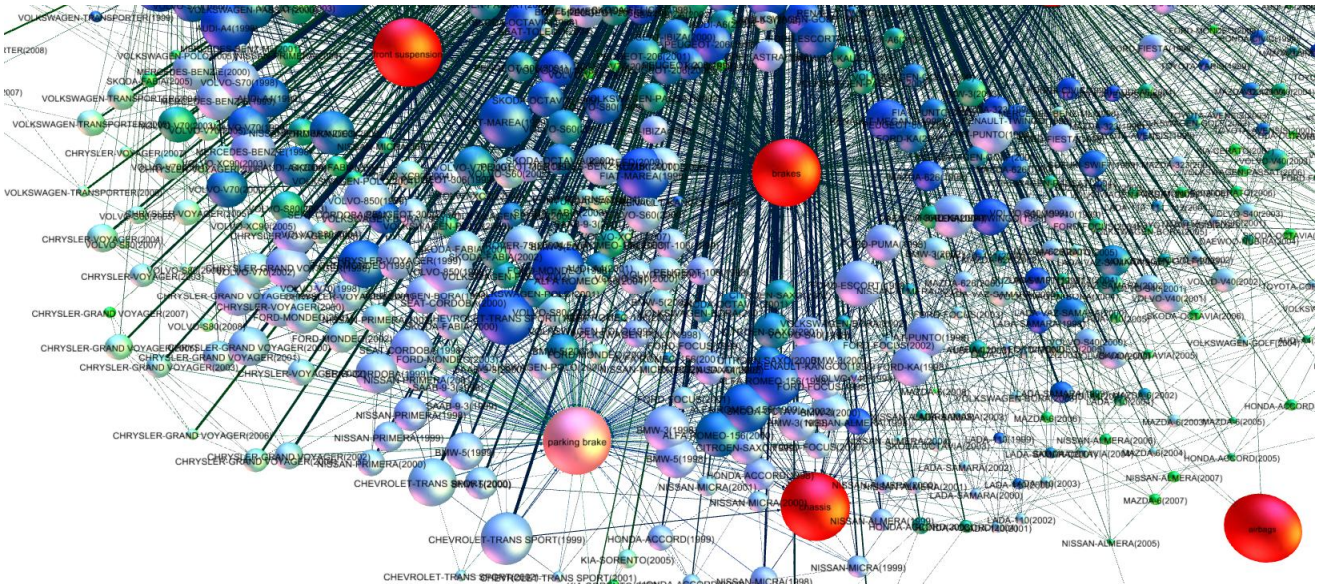


Figure 3. The labels of cars with the parking brake rejection reasons are highlighted.

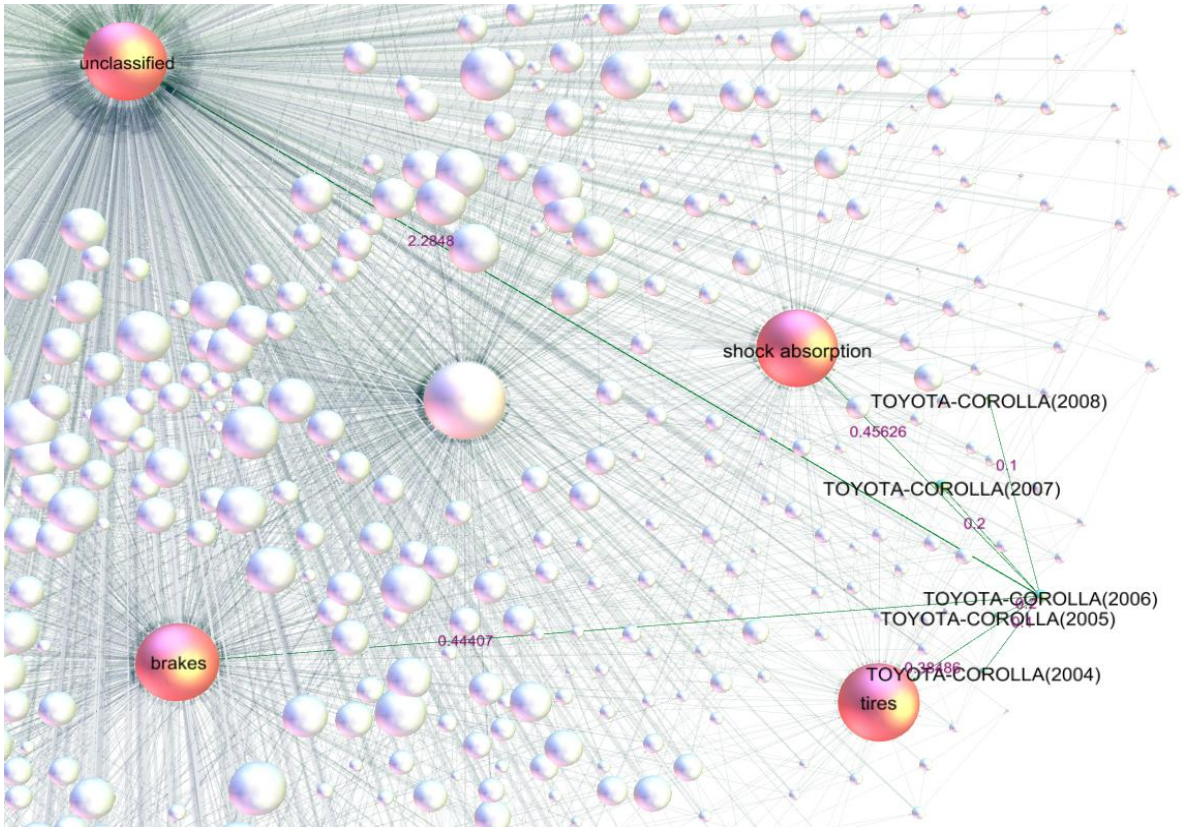


Figure 4. The connections of Toyota Corolla (2006) are shown. It is connected with all rejection reasons mentioned in reports 2009-2011 and the same cars which have been introduced into use between 2004 and 2008. Short edges between Toyota Corolla cars are corresponding similar RRs.

V. CONCLUSION AND FUTURE WORK

Our goal in information visualization was reached even though the provided data was not complete. The achieved results with certain assumptions related to data preprocessing and visualization are reported in this paper. PCA visualization was found to facilitate data exploration to some degree. Exploring the car inspection data is faster using the visualizations in Gephi or in a browser than by the dozens of tables. Visualization with the same parameters as presented in this paper is available on web [10]. A user can study the dependencies between the different rejection reasons and cars by exploring our network visualization.

Our work is still in progress and in future work, we will use additional car inspection data to get more reliable and high quality visualizations. Other layouts will be considered in order to improve the network readability, e.g. [11].

ACKNOWLEDGMENT

We would like to thank Aalto University, Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems (Hecse) and A-Katsastus for providing the data.

REFERENCES

[1] M. Bastian, S. Heymann and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks”, International AAAI Conference on Weblogs and Social Media, 2009.

[2] N.B. Ellison et al., “Social Network Sites: Definition, History, and Scholarship”, Journal of Computer-Mediated Communication, Wiley Online Library, vol. 13, pp. 210–230, 2007.

[3] Discussion Network of the Members in The Finnish Parliament, 2012 http://gexf-js.teelmo.info/index.html#edustajien-puheet-network-2012-party_v2_gt3.gexf, retrieved on June 2012.

[4] J. Heer and D. Boyd, D, “Vizster: Visualizing Online Social Networks”, IEEE Symposium on Inf. Visualization, pp. 32–39, 2005.

[5] TÜV reports http://www.anusedcar.com/, retrieved on June 2012.

[6] J. Talonen and M. Sulkava, “Analyzing Parliamentary Elections Based on Voting Advice Application Data”, Advances of Intelligent Data Analysis, Springer, pp. 340–351, 2011.

[7] http://www.a-katsastus.com/, retrieved on June 2012.

[8] M. Jacomy and T.Venturi, “ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization”, unpublished 2011.

[9] J. Hair, R. Anderson, R. Tatham and W. Black, “Multivariate Data Analysis” Prentice Hall, 5th edition, 1998.

[10] Exported car inspection data network visualization http://dl.dropbox.com/u/7846727/2012gephi/index.html#cars.gexf, retrieved on June 2012.

[11] A.J. Enright and C.A. Ouzounis, “BioLayout – an automatic graph layout algorithm for similarity visualization, Bioinformatics, vol. 17, pp. 853–854, 2001.

 ISBN: 978-1-61208-242-4

Trend Visualization on Twitter: What's Hot and What's Not?

Sandjai Bhulai, Peter Kampstra, Lidewij Kooiman, and Ger Koole
Faculty of Sciences
VU University Amsterdam
Amsterdam, The Netherlands
{*s.bhulai, p.kampstra, ger.koole*}*@vu.nl, l.e.kooiman@gmail.com*

Marijn Deurloo and Bert Kok
CCinq
Amsterdam, The Netherlands
{*marijn, bert*}*@ccinq.com*

***Abstract*—Twitter is a social networking service in which users can create short messages related to a wide variety of subjects. Certain subjects are highlighted by Twitter as the most popular subjects and are known as trending topics. In this paper, we study the visual representation of these trending topics to maximize the information toward the users in the most effective way. For this purpose, we present a new visual representation of the trending topics based on dynamic squarified treemaps. In order to use this visual representation, one needs to determine (preferably forecast) the speed at which tweets on a particular subject are posted and one needs to detect acceleration. Moreover, one needs efficient ways to relate topics to each other when necessary, so that clusters of related trending topics are formed to be more informative about a particular subject. We will outline the methodologies for determining the speed and acceleration, and for clustering. We show that the visualization using dynamic squarified treemaps has many benefits over other visualization techniques.**

***Keywords-microblogging; Twitter; trend detection; clustering; visualization; dynamic squarified treemaps*.**

I. INTRODUCTION

Twitter, a popular microblogging service, has seen a lot of growth since it launched in 2006 and commands more than 140 active million users with 340 million messages (tweets) per day as of March 2012 [1]. Twitter users write tweets about any topic within the 140-character limit and follow others to receive their tweets. An important characteristic of Twitter is its real-time nature. For example, when a major event occurs, people disseminate tweets over the network related to the event, which enables detection of the event promptly by observing the tweets. The popular events and subjects are also known as trending topics, and their detection helps us to better understand what is happening in the world.

The visualization of trending topics is an important research question, since the representation of the trending topics has a significant impact on the interpretation of the topics by the user. This visualization can be done simply by providing a list of topics, as Twitter does (see [2] and Figure 1). However, this representation suffers from a number of drawbacks that prevent the user in assessing the importance of the topic correctly. First, although the list is ordered from the most popular topic to the least popular topic, one cannot infer the importance of each topic relative to the other topics. Second, a list also does not convey the dynamics in the trend, e.g., is the topic still trending to become more popular or is a different topic growing more popular? Third, it could very well be that several topics on the list are related to each other and should be grouped into a coherent set of topics. For example, it is not clear on the outset that topics 3 and 9 in Figure 1 are related to each other. This group of topics could provide more semantics to users than a single topic alone.

1) #PrayforMexico
2) #SocialMovies
3) #temblor
4) Sismo de 7.8
5) Earthquake in Mexico
6) John Elway
7) Pat Bowlen
8) Marcelo Lagos
9) Azcapotzalco
10) Niñas de 13 y 14

Figure 1. Trending topics on Twitter, recorded on 20 March 2012.

A popular method to visualize trending topics is a tag cloud (see Figure 2). However, the research on the effectiveness of this visualization technique is not conclusive. Sometimes, a simple list ordered by frequency may work better in practice than fancy sequential or spatial tag clouds [3]. In other research (e.g., [4]) an alphabetically ordered list performed best with variations in font size (a bigger font for more important topics worked better). Some results show that font size and font weight have stronger visual effects than intensity, number of characters, or tag area. However, when several visual properties are manipulated at once, there is no single property that stands out above the others according to [5]. Hearst and Rosner [6] even argues that "the limited research on the usefulness of tag clouds for understanding information and for other information processing tasks suggests that they are (unsurprisingly) inferior to a more standard alphabetical listing."

A dynamic tag cloud addresses the first two of the three shortcomings of lists to some extent. The importance of each

 ISBN: 978-1-61208-242-4

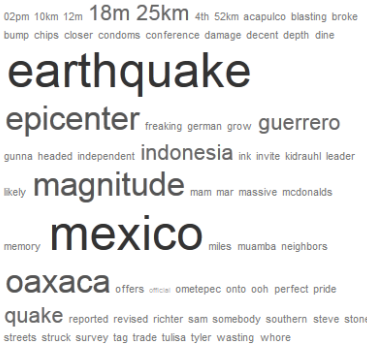


Figure 2. Twitscoop dynamic tag cloud.

topic is displayed by the font size in the tag cloud. The dynamics of the trend can be implemented by a dynamic tag cloud in which the text size grows or shrinks. However, the last shortcoming for addressing topics that are related to each other is more difficult. In this case, one needs to cluster trending topics into coherent groups and visualize them, e.g., through semantics [7], [8]. In order to visualize these clusters, one could use a Treemap [9] or a Squarified Treemap [10], [11]. A treemap displays hierarchical data as a set of nested rectangles.

In this paper, we propose a Dynamic Squarified Treemap (see Figures 6 and 7) to overcome all three aforementioned shortcomings. The importance of a topic can now be correlated to the size of a rectangle. The color of the rectangle can be used to identify if the topic is trending upwards, downwards or remains at its popularity. The rectangle itself can harbor multiple topics so that clusters can be visually represented in an appealing manner. In order to use this visual representation, we need to define how to choose the importance (which is directly related to the number of tweets per second on the topic) and how to choose the color (which is directly related to the acceleration or deceleration of the number of tweets per second).

Our contribution in this paper is threefold. First, we have a different perspective than most other works (e.g., as compared to [11], which is the only paper related to our work). We are focused on upcoming topics that will become a trend instead of a complete online overview of topics. The visualization of these topics is performed dynamically in which color, size, and animation carry additional information. Second, we develop algoritms to quickly determine the importance of topics using new smoothing methods based on little input data. Third, we show that for our purposes simple online clustering techniques perform sufficiently well.

The rest of the paper is structured as follows. In Section II, we outline the methodology to determine the input parameters for the dynamic squarified treemaps. In Section III, we explain how the dynamic aspect of squarified treemaps is more informative than other visualization methods. We conclude the paper with some additional remarks in Section IV.

II. Methodologies

In this section, we outline the methodology to determine the speed of tweets and the acceleration. These two parameters will serve as input parameters for the dynamic squarified treemap to generate a visualization of the trending topics. We first start with the twitter speed of a specific topic. For this purpose, we use the trending topics as posted by Twitter on 20 March 2012; see Figure 1. To illustrate our techniques, we focus on the tweets in hashtag #PrayforMexico. This hashtag was a trending topic at that time as a result of an earthquake in Mexico. The data derived from this hashtag consists of tweets with a time stamp (with seconds as accuracy). Based on this data, the absolute number of tweets over the day is given in Figure 3. One can see that around 7.30pm the number of tweets rapidly increases due to the earthquake.

A. Speed of Tweets

Let us for ease of notation focus on a stream of tweets on a particular subject for which the twitter speed needs to be determined. Let us denote by t_i the time stamp of the i-th tweet with $t_1 \leq t_2 \leq \cdots$. The speed can in principle be determined by a simple moving average, e.g., when tweet i arrives, the speed v_i can be determined by $k/(t_i - t_{i-k})$ for some k that determines how much history is included. There are two significant drawbacks to such a method. First, for high volume tweets (in particular, for popular topics), many tweets have the same time stamp. Thus, it could be that $t_i = \cdots = t_{i-k}$ so that v_i is not well-defined due to division by zero. Second, such an approach looks back at the history and has little predictive power.

To alleviate the drawback of the moving average, we first determine the interarrival times $a_i = t_i - t_{i-1}$. When tweet i is recorded, it could be that there are already several tweets that have the same time stamp (this is the case when $a_i = 0$). This number is given by $z_i = |\{k \,|\, t_i = t_k\}|$. Hence, we adjust the time stamp of the tweets by spreading them uniformly over the past second. Thus, we transform a_i to a'_i by

$$a'_i = \begin{cases} a_i - \left(1 - \frac{1}{z_i+1} - \frac{1}{z_{i-1}+1}\right), & a_i > 0, \\ \frac{1}{z_i+1}, & a_i = 0. \end{cases}$$

Next, we apply exponential smoothing with parameter $0 \leq \alpha \leq 1$ on the new interarrival times to derive a new time series b_i given by

$$b_i = \alpha b_{i-1} + (1-\alpha) a'_i,$$

starting with $b_1 = a'_1$. Since the resulting time series can still be too volatile, we apply a double smoothing by taking the average over the past k values of the time series b_i. Thus, the speed v_i (in tweets per second) is then given by

$$v_i = \frac{k}{\sum_{j=i-k+1}^{i} b_j}.$$

 ISBN: 978-1-61208-242-4

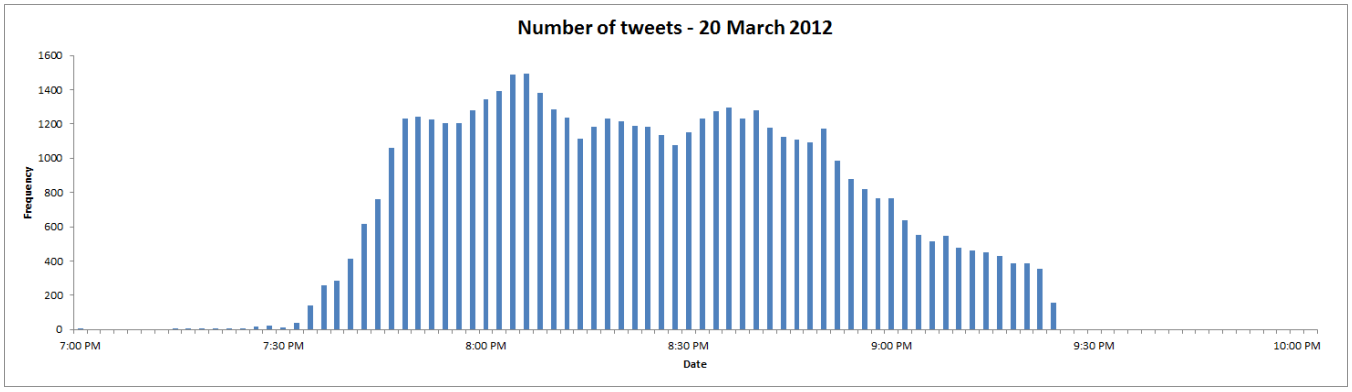


Figure 3. Absolute number of tweets for #PrayforMexico on 20 March 2012.

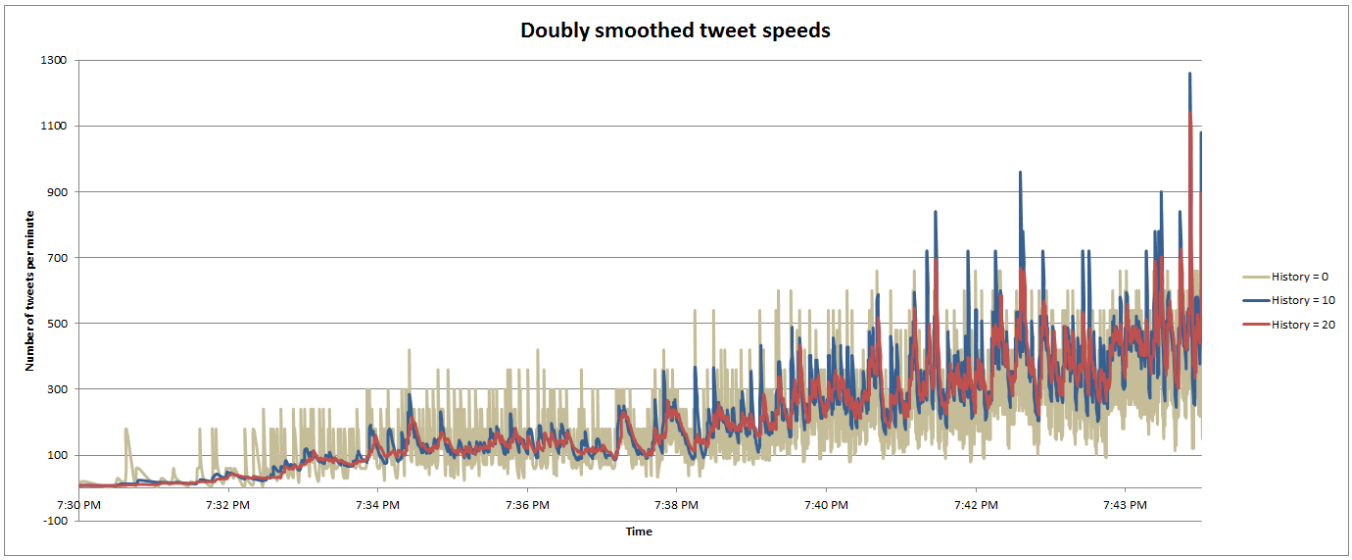


Figure 4. The number of tweets per second for different values of k (history) for #PrayforMexico on 20 March 2012.

Our algorithm thus has two parameters that can be chosen freely. We have the first smoothing parameter α that is used in the exponential smoothing, and we have the second smoothing parameter k that uses k tweets from history. In Figure 4, we can see the graph of the tweets from #PrayforMexico for various values of k. The parameter α is set to 0.8, which seems to work best for various examples in our setting. We can see that a value of $k = 0$, the case in which no history is taken into account, is rather volatile and does not produce stable results. The values of $k = 10$ and $k = 20$ provide more stable results and are much smoother than the graph for $k = 0$.

B. Acceleration of Tweets

The acceleration of tweets is basically a derivative of the speed of the tweets. We calculate the acceleration of the tweets for each minute. Let t be the start of a minute and $t+1$ the start of the next minute. Let $\overline{z}_t$ be the index of the last tweet before the end of the minute, thus $\overline{z}_t = \max\{i \mid t_i < t+1\}$. Denote by $\underline{z}_t$ the first tweet in that minute, or if there are not any, the one before that. Thus $\underline{z}_t = \max\{\min\{i \mid t \leq t_i < t_{\overline{z}_t}\}, \overline{z}_t - 1\}$. The acceleration w_t is then computed by

$$w_t = \frac{v_{\overline{z}_t} - v_{\underline{z}_t}}{t_{\overline{z}_t} - t_{\underline{z}_t}}.$$

Note that the definition closely reflects the regular definition of a derivative. However, we account for the fact that there can be no tweets in a particular minute. This is taken care by the way the variables $\overline{z}_t$ and $\underline{z}_t$ are defined. Furthermore, we also account for the fact that all tweets in the minute can have the same time stamp. Therefore, we use the adjusted timestamps a'_i instead of a_i.

In Figure 5, we can see the graph of the acceleration of the number of tweets per second for different values of k (the history that is used to determine v_i). As in the case of the calculation of the speed, we conclude that $k = 0$ (not using any history at all) results in volatile accelerations that are not preferred. Since the graph of speeds when using $k = 10$ is still very bursty, the acceleration shows large fluctuations that are not in accordance with ones intuition

 ISBN: 978-1-61208-242-4

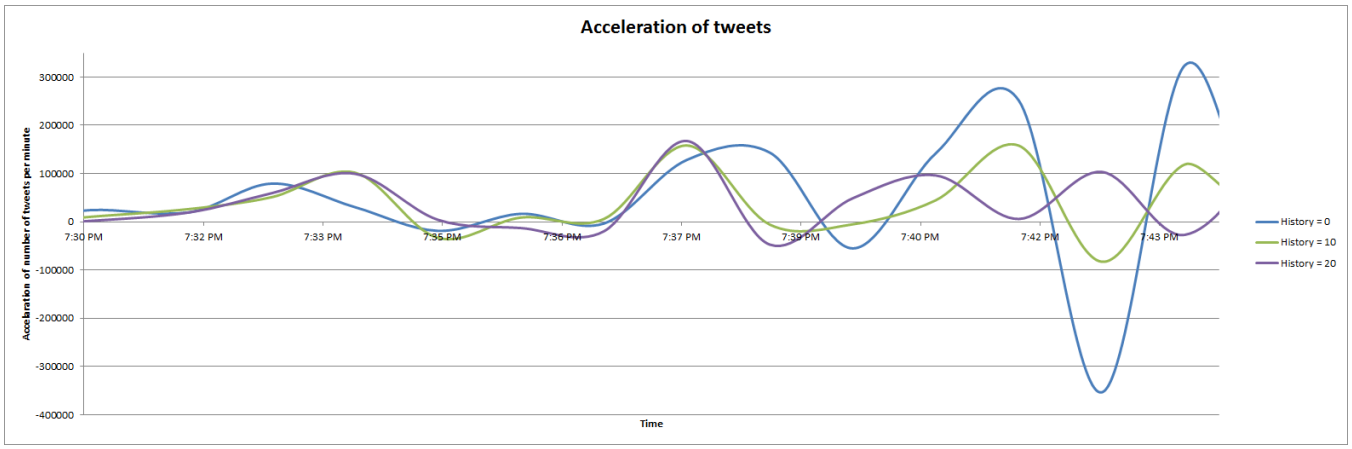
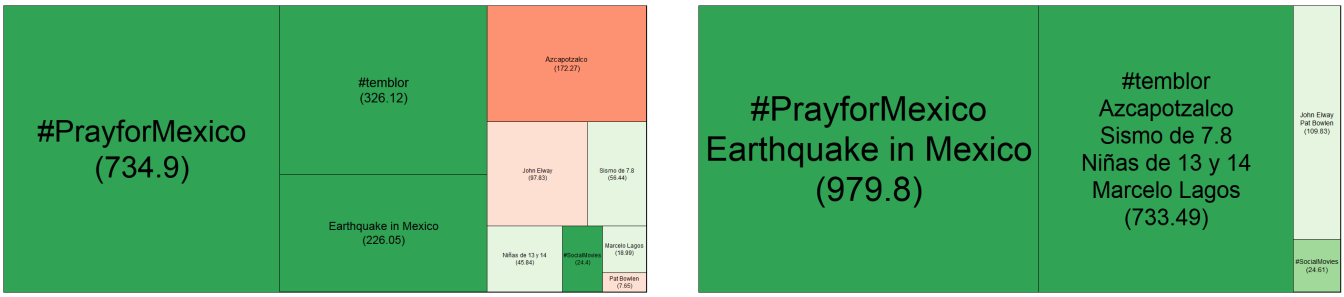


Figure 5. The acceleration of the number of tweets per second for different values of k (history) for #PrayforMexico on 20 March 2012.



(a) Unclustered (b) Clustered

Figure 6. Numerical results.

(see, e.g., the timestamps around 7.42pm). The graph with $k = 20$, however, seems to perform well in this case, and in other cases as well. We can clearly see that the acceleration is picked up at 7.32pm, which corresponds to a real surge in the absolute number of tweets. Thus, this is precisely the moment at which one would like to detect this trend. Hence, in the rest of the paper, our algorithms run with $\alpha = 0.8$ and $k = 20$.

III. Dynamic Squarified Treemaps

In the previous section, we have identified the major ingredients for building a squarified treemap. First, we have determined the variable v_i, which represents the twitter speed in tweets per second on a particular topic. Second, we have identified the acceleration w_t of the number of tweets for the same topic. Based on this information, we build rectangles for each topic of which the relative areas correspond to the relative speeds of each topic. On top of that, each rectangle is color-coded from green to white to red, based on a positive to neutral to negative acceleration. This gives rise to a representation as depicted in Figure 6a. The numbers in parentheses represent v_i. This representation solves many of the issues tied to lists (see Figure 1) and tag clouds (see Figure 2). In this section, we improve the visual representation by clustering related topics.

A. Clustering topics

The clustering of tweets is not an easy process. Standard algorithms, such as K-means clustering [12], are slow. Therefore, most algorithms usually work iteratively. For speed, a single assignment is usually used in the literature (e.g., [13], [14]).

A simple way to cluster tweets is by using a cosine similarity as defined in [15]. In this algorithm, the term frequency and inverse document frequency (TF-IDF) [16] can be used as a weighing scheme. A more involved method to cluster tweets is the Latent Dirichlet Allocation (LDA) [17], which can be used to track topics over time [18]. The clustering that is obtained by this method is better than when using TF-IDF [19] (while a combination works best). However, LDA is not perfect for Twitter because tweets are limited in size [20]. Methods based on non-negative Matrix Factorization [21] could be an alternative to TF-IDF and LDA (from [22]). Some experimentation has already been performed in [23] on a small dataset. One can also think of mixture models [24], [25], which were developed for producing recommendations, for clustering tweets.

B. Clustering based on tweet list comparison

As a first clustering algorithm, we adopt a very simple but efficient clustering algorithm. For each topic a, at time

 ISBN: 978-1-61208-242-4

Table I
CLUSTERING BASED ON COMPARISON OF TWEET LISTS.

	#PrayforMexico	#temblor	Earthquake in Mexico	Azcapotzalco	John Elway	Sismo de 7.8	Niñas de 13 y 14	#SocialMovies	Marcelo Lagos	Pat Bowlen
#PrayforMexico	1	0.01	0.18	0.01	0.01	0.04	0	0.01	0.01	0
#temblor	0.01	1	0.02	0.02	0.01	0.01	0	0.01	0.01	0
Earthquake in Mexico	0.18	0.02	1	0.01	0.01	0.03	0	0.01	0.01	0
Azcapotzalco	0.01	0.02	0.01	1	0.01	0.01	0	0.01	0.01	0
John Elway	0.01	0.01	0.01	0.01	1	0.01	0	0.01	0.01	0.04
Sismo de 7.8	0.04	0.01	0.03	0.01	0.01	1	0	0.01	0.01	0
Niñas de 13 y 14	0	0	0	0	0	0	1	0	0	0
#SocialMovies	0.01	0.01	0.01	0.01	0.01	0.01	0	1	0.01	0
Marcelo Lagos	0.01	0.01	0.01	0.01	0.01	0.01	0	0.01	1	0
Pat Bowlen	0	0	0	0	0.04	0	0	0	0	1

Table II
CLUSTERING BASED ON THE COSINE SIMILARITY INDEX.

	#PrayforMexico	#temblor	Earthquake in Mexico	Azcapotzalco	John Elway	Sismo de 7.8	Niñas de 13 y 14	#SocialMovies	Marcelo Lagos	Pat Bowlen
#PrayforMexico	1	0.30	0.70	0.21	0.14	0.27	0.19	0.16	0.21	0.13
#temblor	0.30	1	0.20	0.47	0.11	0.54	0.41	0.12	0.42	0.10
Earthquake in Mexico	0.70	0.20	1	0.15	0.16	0.17	0.11	0.16	0.14	0.15
Azcapotzalco	0.21	0.47	0.15	1	0.08	0.48	0.33	0.09	0.33	0.08
John Elway	0.14	0.11	0.16	0.08	1	0.06	0.06	0.11	0.08	0.32
Sismo de 7.8	0.27	0.54	0.17	0.48	0.06	1	0.46	0.08	0.38	0.06
Niñas de 13 y 14	0.19	0.41	0.11	0.33	0.06	0.46	1	0.06	0.33	0.06
#SocialMovies	0.16	0.12	0.16	0.09	0.11	0.08	0.06	1	0.10	0.11
Marcelo Lagos	0.21	0.42	0.14	0.33	0.08	0.38	0.33	0.10	1	0.07
Pat Bowlen	0.13	0.10	0.15	0.08	0.32	0.06	0.06	0.11	0.07	1

t, we keep a list l_a of the last 100 tweets counting back from time t. Our similarity metric for topic a and topic b is defined as the number of times that both terms a and b appear in the lists l_a and l_b. If the similarity metric is above the threshold of 0.15, then the two topics are clustered, and clustering continues until no more tokens can be added to the cluster. Table I displays the results of this clustering. In the results we can see that '#PrayforMexico' and 'Earthquake in Mexico' are clustered.

C. Clustering based on the cosine similarity index

We also adopt the cosine similarity [26] to cluster the tweets. The cosine similarity of two topics a and b is a measure of similarity, defined by

$$\frac{< f_a, f_b >}{||f_a|| \cdot ||f_b||} = \frac{\sum_i f_a(i) f_b(i)}{\sqrt{\sum_i f_a(i)^2} \sqrt{\sum_i f_b(i)^2}},$$

where the vector f_a (and f_b) is the frequency list of terms that appear in the list l_a (and l_b). The cosine similarity is bounded between 0 and 1 since both f_a and f_b are non-negative. The name of the similarity index is derived from the interpretation of the cosine of the angle between the two vectors. Hence, similar vectors (with an angle close to zero) have a high cosine similarity, whereas vectors that are not similar (with an angle close to $\pi/2$) have a low cosine similarity. If the similarity metric is above the threshold of 0.30, then the two topics are clustered. Table II displays the results of this clustering. In the results we can see that '#PrayforMexico' and 'Earthquake in Mexico' are in one cluster. In addition, '#temblor', 'Azcapotzalco', 'Sismo de 7.8', 'Niñas de 13 y 14', and 'Marcelo Lagos' form one cluster, as well as 'John Elway' and 'Pat Bowlen'. Observe that the two largest clusters are actually about the same subject, but in two different languages. A human observer would either put these into one cluster or into two. In fact, our clustering algorithm almost puts these into one cluster, with a cosine similarity of 0.30.

Figure 6b shows the squarified treemap for the clustered topics. It is clear that this representation is even better than Figure 6a. From the clusters is becomes clear that 'Azcapotzalco' is related to the earthquake in Mexico, although this was not clear before. Figure 7 depicts the dynamic part of the squarified treemaps. Using jQuery [27], the tiles in the treemap transition to their new size and position based on the newly calculated speed and acceleration values. This dynamic part has the appealing feature that one can directly identify visually the emerging and receding topic. The dynamic clustered squarified treemap resolves the three issues that were mentioned as problems with lists and tag clouds. Experiments with test persons seem to suggest that the dynamic squarified treemap is an effective method for the display of dynamic data from Twitter.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have discussed the dynamic squarified treemap for visually representing the trending topics on Twitter. The main ingredients for this graph are the speed of tweets and the acceleration of them. We have developed algorithms to calculate both of them. Moreover, we have discussed a simple clustering algorithm to deal with grouping related topics in online twitter streams. The final representation in a dynamic squarified treemap fills the gaps

 ISBN: 978-1-61208-242-4

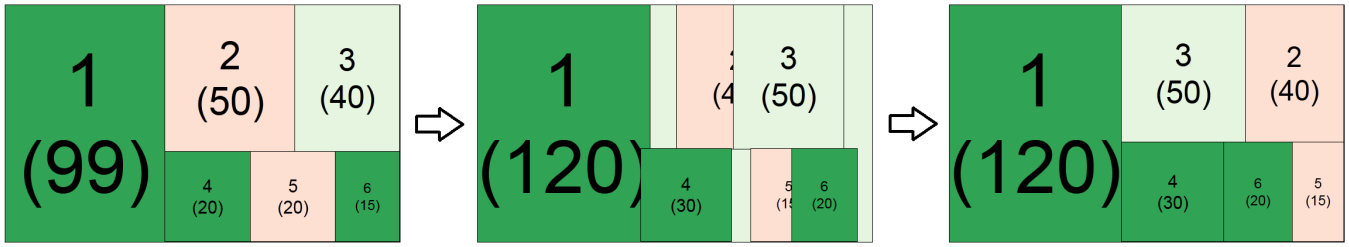


Figure 7. The transitions in the dynamic squarified treemap.

that are present in list and tag cloud representations. Hence, the dynamic squarified treemap forms a powerful visual tool to visualize trending topics.

The analysis in this paper has been done on the trending topics based on the list provided by Twitter. However, we are currently working on a system in which we monitor a sample of the twitter stream and detect trending topics ourselves. The system calculates the speed and acceleration every second and updates the screen accordingly. Based on the size and rate of growth of a cluster of words / topics the dynamic squarified treemap serves as an early warning system for trends.

REFERENCES

[1] Wikipedia, "Twitter," URL: en.wikipedia.org/wiki/Twitter.

[2] Twitter, URL: www.twitter.com.

[3] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting our head in the clouds: toward evaluation studies of tagclouds," in *Proc. of the SIGCHI Conf. on Human factors in computing systems*, 2007, pp. 995–998.

[4] M. Halvey and M. Keane, "An assessment of tag presentation techniques," in *Proc. of the 16th Intl. Conf. on World Wide Web*. ACM, 2007, pp. 1313–1314.

[5] S. Bateman, C. Gutwin, and M. Nacenta, "Seeing things in the clouds: the effect of visual features on tag cloud selections," in *Proc. of the 19th ACM Conf. on Hypertext and hypermedia*, New York, NY, USA, 2008, pp. 193–202.

[6] M. Hearst and D. Rosner, "Tag clouds: Data analysis tool or social signaller?" in *Hawaii Intl. Conf. on System Sciences, Proc. of the 41st Annual*. IEEE, 2008, pp. 160–160.

[7] L. Di Caro, K. S. Candan, and M. L. Sapino, "Using tagflake for condensing navigable tag hierarchies from tag clouds," in *Proc. of the 14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008, pp. 1069–1072.

[8] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, "Semantic grounding of tag relatedness in social bookmarking systems," in *Proc. of the 7th Intl. Conf. on The Semantic Web*, ser. ISWC '08. Berlin: Springer-Verlag, 2008, pp. 615–631.

[9] B. Shneiderman and M. Wattenberg, "Ordered treemap layouts," in *Proc. of the IEEE Symp. on Information Visualization 2001 (INFOVIS'01)*, Washington, DC, USA, 2001, pp. 73–.

[10] M. Bruls, K. Huizing, and J. Van Wijk, "Squarified treemaps," in *Proc. of the Joint Eurographics and IEEE TCVG Symp. on Visualization*. Citeseer, 2000, pp. 33–42.

[11] D. Archambault, D. Greene, P. Cunningham, and N. Hurley, "Themecrowds: multiresolution summaries of twitter usage," in *Proc. of the 3rd Intl. Workshop on Search and mining user-generated contents*. ACM, 2011, pp. 77–84.

[12] A. Karandikar, "Clustering short status messages: A topic model based approach," Master's thesis, Faculty of the Graduate School of the University of Maryland, 2010.

[13] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in *Proc. of the 5th Intl. AAAI Conf. on Weblogs and Social Media*, 2011.

[14] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proc. of the 17th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems*, 2009, pp. 42–51.

[15] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proc. of the 27th Annual Intl. ACM SIGIR Conf. on Research and development in information retrieval*, 2004, pp. 297–304.

[16] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Mngmnt*, vol. 24, pp. 513–523, August 1988.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[18] D. Knights, M. C. Mozer, and N. Nicolov, "Detecting topic drift with compound topic models," in *Proc. of the Fourth Intl. AAAI Conf. on Weblogs and Social Media*, 2009.

[19] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proc. of the Fourth Intl. AAAI Conf. on Weblogs and Social Media*, 2010.

[20] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," in *Proc. of the ACM SIGIR 3rd Workshop on Social Web Search and Mining*, 2011.

[21] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct 1999.

[22] S. P. Kasiviswanathan, P. Melville, A. Banerjee, and V. Sindhwani, "Emerging topic detection using dictionary learning," in *20th ACM Conf. on Info. and Knowledge Mngmnt*, 2011.

[23] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization," in *Proc. of the Fifth ACM Intl. Conf. on Web search and data mining*, 2012, pp. 693–702.

[24] J. Kleinberg and M. Sandler, "Using mixture models for collaborative filtering," in *Proc. of the Thirty-Sixth Annual ACM Symp. on Theory of computing*, 2004, pp. 569–578.

[25] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in *Proc. of the Tenth ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, New York, NY, USA, 2004, pp. 811–816.

[26] P.-N. Tan, M. Steinbach, and V. Kumar. "Introduction to Data Mining," Addison-Wesley, 2005, Chapter 8.

[27] jQuery, URL: jquery.com.

 ISBN: 978-1-61208-242-4

Travel Time Estimation Results with Supervised Non-parametric Machine Learning Algorithms

Ivana Cavar, Zvonko Kavran, Ruder Michael Rapajic
Faculty of transport and traffic sciences
University of Zagreb
Zagreb, Croatia
ivana.cavar@fpz.hr; zvonko.kavran@fpz.hr; miso.rapajic@fpz.hr

***Abstract*— Paper describes urban travel time estimation procedure based on non-parametric machine learning algorithms and three data sources (GPS vehicle tracks, meteorological data and road infrastructure data base). After data fusion and dimensionality reduction, new road classification is defined and for four different time intervals and five different road categories travel time estimation is conducted. For travel time estimation, *k* nearest neighbors (kNN) and IRM-based (Iterative Regression Method) approaches were applied. Best results for two hour forecasting period are achieved for road class with highest traffic flow.**

Keywords-GPS vehicle track; travel time estimation; k-nearest neighbours; iterative regression model; urban traffic.

I. INTRODUCTION

Travel time estimate is important information applicable in many intelligent transportation system (ITS) services (route guidance, advanced traveler information system or advanced traffic management system) as well as in transportation planning process. It is one of the largest costs of transportation, and travel time savings are often the primary justification for transportation infrastructure improvements. Travel time information is acceptable and useful to all, decision makers, transport system users, technical and nontechnical staff and often used as relevant when comparing different transportation modes.

Today's accessibility of different data sources and large quantity of data has double effect on travel time estimation. It eases up the procedure by allowing different and up to date quality data to be included in estimation. On the other hand it goes under the 'curse of dimensionality' by making calculation process more demanding if even possible. Therefore statistical methods are not applicable in these cases and advanced analytics and data science mechanisms are required.

In literature, different approaches for travel time estimation can be found. Although there are papers dealing with roads that have different characteristics [1][2], best results are achieved for freeways and free flow traffic conditions [5].Reason for this lies in the smaller number of factors influencing the travel time.

Understanding traffic factors affecting travel time is essential for improving prediction accuracy. Some of traffic factors that affect travel time are free flow travel speed [1], occurrence of incident situations, holidays or other uncommon events [2], congestion level [3] and weather conditions [4][5].

Another important element of travel time estimation is the forecasting period (the greater it is the higher is the prediction error) [6].

When it comes to travel time estimation approaches applied in literature, they can be divided in two basic groups [7]:

- Extrapolation models - mainly based on statistical approaches and historical values as regression models [8] [9], ARIMA (Autoregressive Integrated Moving Average) models [20], STARIMA (Space-Time Autoregressive Integrated Moving Average) [10], Kalman filter [11], ANN (artificial neural networks) [12], SVM Support vector machines [13] and pattern based forecasting [14].
- Explanatory models - based mainly of factor or parameters analyses and traffic flow theory as dynamic traffic assignment [15] [16].

Paper has six sections. After introduction, a description of travel time estimation methodology is given. A case study and result comparison are defined, followed by a conclusion.

II. DESCRIPTION OF TRAVEL TIME ESTIMATION METHODOLOGY

Two non-parametric methods, kNN and IRM, are used for travel time estimation and described in more details.

A. K nearest neighbours

K nearest neighbors is an algorithm that is based on similarity metric described by distance function locates the "best" neighbors (the neighbors that are most likely similar to the target). One of the most popular choices to measure this distance is Euclidean. Other measures include Euclidean squared, City-block, and Chebyshev [17]:

$$D(x,p) = \begin{Bmatrix} \sqrt{(x-p)^2} & Euclidean \\ (x-p)^2 & Euclidean\ squared \\ Abs\ (x-p) & Cityblock \\ Max\ (|x-p|) & Chebyshev \end{Bmatrix} \quad (1)$$

where *x* and *p* are the query point and a case from the examples sample, respectively.

 ISBN: 978-1-61208-242-4

The choice of k can be regarded as one of the most important factors of the model that can strongly influence the quality of predicted outcome. The goal is to find an optimal value for k that achieves the right trade-off between the bias and the variance of the model. For travel time estimation purpose, value of k for each road class was determined by cross validation method.

For regression, a kNN prediction is the average of the k-nearest neighbor outcome.

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (2)$$

where y_i is the ith case of the examples sample and y is the prediction (outcome) of the query point. For classification problems, like road classification, kNN predictions are based on a voting scheme in which the winner is used to label the query.

B. *Iterative regression method*

IRM is iterative multivariate linear regression approach where estimation for dependent continuous variable is based on determination of unknown number of independent variables from set of all proposed predictors.

Estimation procedure is described through following steps:

1. Defining dependent variable y and $z_1, \ldots, z_n$ independent variables.
2. Calculate correlation coefficient between y and each of independent variables $z_1, \ldots, z_n$.
3. Select independent variable with the most significant correlation coefficient.
4. Mark selected independent variable with x_1 and calculate parameters a_1 and b_1 as first iteration by linear regression function:
$$y_1 = a_1 + b_1 x_1 \quad (3)$$
5. Determine the residual r:
$$r_1 = y_1 - (a_1 + b_1 x_1) \quad (4)$$
6. Calculate correlation coefficient between residual (r_1) and each of remaining independent variables.
7. Select independent variable with the most significant correlation coefficient.
8. Mark selected independent variable with x_2 and calculate parameters a_2 and b_2 as second iteration by linear regression function:
$$r_2 = a_2 + b_2 x_2 \quad (5)$$
9. Regression functions from first and second iteration give more precise estimation of dependent variable:
$$y_1 = (a_1 + a_2) + b_1 x_1 + b_2 x_2 \quad (6)$$
10. Repeat steps 5 to 9 until residual r becomes small enough (stops effecting 0,001% of predicted value for y in previous iteration)

C. *Applied travel time estimation procedure*

Methodology proposed for travel time estimation procedure incorporates multiple data analytics techniques. Overall overview of used methods and procedures is given on Figure 1.

First phase incorporates different data collection methods, data fusion and cleansing, as well as data interpolation (when needed). Product of these procedures is input for time series data analysis that results with detection of characteristic time intervals. This is done based on the analysis of speed records. Characteristic time intervals are considered to be time periods in which the vehicle throughput capacity on the observed part of the road is significant. For example, significantly low (to identify the morning and afternoon "peak" periods) or significantly high (free flow conditions).

After this phase, gathered data are used for revision of official road classification. Reason for this lays in the fact that official road classifications are mainly used for general purposes and roads are classified based on infrastructural characteristics. For deep traffic analysis like travel time estimation this classification is not sufficient and new classification is developed. In this paper, classification methodology is used to classify road segments (continuous part of the road from one intersection to the sequent one) based on real traffic flow data collected from GPS tracks as

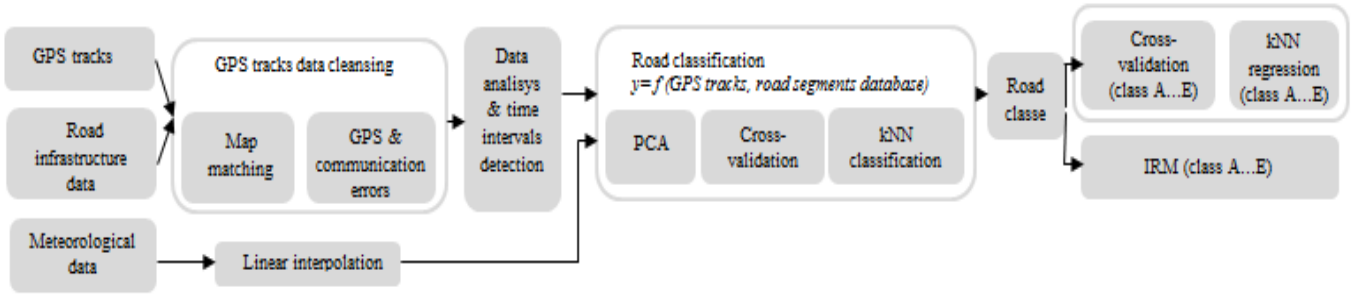


Figure 1. Travel time estimation procedure.

well as road segments infrastructure data. The software used for this procedure is StatSoft Statistica Data Miner. Priori the application of classification technique called kNN principle component analysis (PCA) was applied. The purpose of this was to reduce dimensionality of collected data and to select variables relevant for road classification.

Since kNN algorithm is highly sensitive to the choice of *k*, after the reduction of dimensionality a *v* fold cross-validation procedure is applied to determine k value. Distance measure used in kNN classification is Euclidian distance for standardized data. Data are standardized to transform all values (regardless of their distributions and original units of measurement) to compatible units from a distribution with a mean of 0 and a standard deviation of 1. This makes the distributions of values easy to compare across variables and/or subsets and makes the results entirely independent of the ranges of values or the units of measurements. Results of this phase are road segments separated in different categories. For each of these categories, in next step, different travel time estimation procedure is carried out.

At this point, whole data set is divided into two subsets, one for training and one for model development. Travel time estimation is based on two approaches kNN regression in combination with v fold cross-validation and IRM approach.This is applied separately for each road class and each characteristic time interval.

D. Advantages and areas of practice

Developed travel time estimation procedure has multiple advantages for practical application as:

- Supports application of multiple data sources (different sensors);
- Allows automated data fusion and dimensionality reduction;
- Improves estimation procedure by basing estimation steps on real traffic performance data (road classification is based on traffic data, not only on road infrastructure data);
- Once defined values of *k* for each road class (in cross-validation step) doesn't need to be calculated again. Due to the fact that this is the most computational demanding step in travel time estimation procedure, future application of this procedure is much less time and computational expensive and therefore suitable for real time travel estimation:
- Completely automated travel time estimation that can give results in real time.

Travel time is one of most important values in traffic planning and allows decision maker to base its decisions on traffic performance on different routes/modes choice etc. Travel time estimation information can be delivered to drivers to aim them in route selection choice or to be a part of automated route guidance system as well as to be incorporated in different ITS services.

III. TRAVEL TIME ESTIMATION (CASE STUDY: CITY OF ZAGREB)

The travel time estimation procedure we propose in this paper is applied for case study in City of Zagreb and described through description of data collection process, road classification, travel time estimation and evaluation of results.

A. Data collection process

Data used for travel time estimation were collected from three sources including 51 835 560 GPS tracks, road segments data and meteorological data from official Meteorological and Hydrological Service. Geographical area covered with data collection process was urban area of City of Zagreb, Croatia and time window for data collection was thirteen months.

B. Data on the traffic conditions

Traffic conditions are analyzed based on the GPS vehicle tracks. The data were recorded by the device and sent via mobile network. Data were sent to the local server and stored in the database (Table 1).

TABLE 1. TRAFFIC DATA

Recorded data	Variable description
Log time	time of recording expressed in UTC (Universal Time Coordinated) standard
Vehicle ID	identifier of the observed vehicle / GPS device
X coordinate	x coordinate of the GPS record (WGS84 – World Geodetic System 1984)
Y coordinate	y coordinate of the GPS record (WGS84)
Speed	current speed in [km/h]
Course	angle at which the vehicle is travelling with reference to the North
GPS status	indicates the accuracy of the record. GPS status 3 indicates that the data have been collected from 4 or more satellites and GPS status 1 that data have been collected from fewer than 2 satellites.
Engine status	shows whether the vehicle engine was running or was turned off while making the recording.

C. Data on the road infrastructure

Road infrastructure data describe road segments and are stored in the form of digital map database (direction, blocked turns, length, marked pedestrian zones, etc.). The elements included in this database are presented in Table 2.

 ISBN: 978-1-61208-242-4

TABLE 2. ROAD INFRASTRACTURE DATA

Record	Variable description
Segment *ID*	identifier of the road segment
Type	numeric code of the road type according to official classification defined in Spatial plan of City of Zagreb
Direction	code representing road direction
Start x	x coordinate of the beginning of segment (WGS84)
Start y	y coordinate of the beginning of segment (WGS84)
End x	x coordinate of the end of segment (WGS84)
End y	y coordinate of the end of segment (WGS84)
Length	length of the segment in [m]
Name	name of road which contains the respective segment

Presentation of road segment data in a form of a digital map is given in Figure 2.

Figure 2. Part of a digital map representing geographic area of City of Zagreb, Croatia.

D. Meteorological data

Meteorological data are received from Meteorological and Hydrological Service. This database includes variables described in Table 3.

TABLE 3. METEOROLOGICAL DATA

Meteorological data
Minimum and maximum daily air temperature and their difference;
Daily air temperature
Wet ground temperature and ground condition
Air pressure and air steam pressure
Sun shining period;
Precipitation and relative humidity;
Snow cover thickness and thickness of new snow cover;
Horizontal visibility
Direction and speed of wind
Ground temperature at -2cm, -5cm, -10cm, -20cm, -30cm, -50cm, -100cm
Observatory diary data

More on data collection and cleansing procedure can be found in literature [18].

E. Road classification

After the collection and fusion of gathered data and priori the development of travel time estimation, road classification has been revised (official road classification in City of Zagreb is defined by Spatial plan of City of Zagreb [19]).

As a result of the procedure proposed on Figure 1, in Road classification box, a set of four distinctive variables is selected:

- mean average speed on road segment,
- speed's standard deviation on road segment,
- length of road segment,
- vehicle count of each road segment.

Based on this, for road classification 28NN model [5] is used (*k*-nearest neighbors model that for segment classification takes into account neighborhood of 28 most similar road segments). Results of classification procedure separates roads in five classes marked as class A, class B, class C, class D and class E. Compared to official road classification that contains also five road classes (highways, fast roads, state, county and local roads) that are divided based on their construction characteristics and the level of authority in charge of their maintenance this way achieved results are distinct in 47.22 % cases.

More on extracted road classes is given in Table 4, where number and share of segments included in each class, overall length of segments and their share in total road network length, number of records/segment length, mean average of number of records on each segment, vehicle speed mean average value, vehicle speed standard deviation and average segment length for each class are presented. Based on this data we can see that class E includes largest number of segments (around 64 %) and larges share (although somewhat lower than segment count share). By comparison of length and segments shares it is visible that class A contains smallest amount of segments but they have lower number of intersections and higher lengths then segments in others classes (length share/segments share for class A equals 3.07). Also, 2.96% of all segments belong to the class A, 11.76% to classes A and B and 30.36% to classes A, B and C. These classes have highest traffic flow and 94.7% of all records were recorded there. Based on average number of records for each class we can see that number of records per segment length decreases from road class A to road class E in non-linear manner. Mean average of number of records on each segment, vehicle mean average speed and its standard deviation as well as average length of road segment decreases from road class A to road class E. And while average speed for class A represents double average speed of class E deviation is increased. We can say that road class A while having highest traffic flow and average vehicle speed also has largest discontinuity in vehicle movements. Respectively, class E has lowest vehicle speed deviation and most continuous traffic flow.

More on road classification based on hybrid approach can be found in literature [5].

 ISBN: 978-1-61208-242-4

TABLE 4. ROAD SEGMENTS DATA

Raoad class	Number of segments	Segments share [%]	Length of road segments in class [km]	Length share [%]	Number of record / segment length [m]	Mean average of number of records on each segment	Vehicle speed mean average value	Vehicle speed standard deviation	Average segment length [m]
Class A	392	2.96	114 210	9.08	26	7664	62.33	15.91	292.1
Class B	1 166	8.8	175 613	13.96	15.7	2357	49.68	14.18	150.6
Class C	2 463	18.6	314 139	24.97	8.8	1123	42.19	12.28	127.5
Class D	714	5.39	67 151	5.34	4.1	386	41.52	11.49	94.1
Class E	8 508	64.25	586 902	46.65	3.5	240	31	10.01	69

IV. TRAVEL TIME ESTIMATION

The travel time estimation procedure is described through definition of characteristic time intervals and application of kNN and IRM methods.

A. *Definition of characteristic time intervals for travel time estimation*

For the purpose of travel time estimation characteristic, time intervals should be identified. A presentation of one such analysis is given in Figure 3 for segment Slavonska Avenue, direction East-to-West. Figure 3 clearly shows the morning congestion which is most expressed in the interval between 06:00 h and 08:00 h and the afternoon congestion which is a bit longer, taking from 15:30 h -18:30 h.

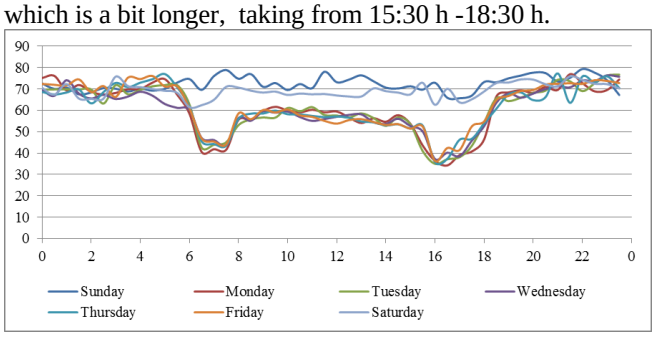


Figure 3. Segment speed analysis

Based on such analysis characteristic time intervals for travel time estimation are defined as follows:

- Morning peak interval (06:00-08:00 h);
- Noon free flow interval (12:00-13:00 h);
- Afternoon peak interval (15:30-18:30 h);
- Evening free flow interval (21:00 – 22:00 h).

B. *Urban travel time estimation procedure*

At this point, whole data set is divided into two subsets. Data collected during first eleven months represents training set for model development. Data collected during last two months are removed from model development procedure and will serve for the travel time estimation error testing.

C. *KNN based approach*

For each road class one road, belonging to that class, is selected to serve as demonstrative one for results of travel time estimation procedure for selected time intervals. These roads are defined in Table 5.

TABLE 5. SHOW UP ROADS FOR EACH ROADS CLASS

Road class	Demonstrative road
A	Slavonska avenue
B	Selska road
C	Savska road
D	Prisavlje
E	Jordanovac

Results of *v* fold cross-validation are presented on Figures 4, 5, 6, 7 and 8, as well as in Table 5.

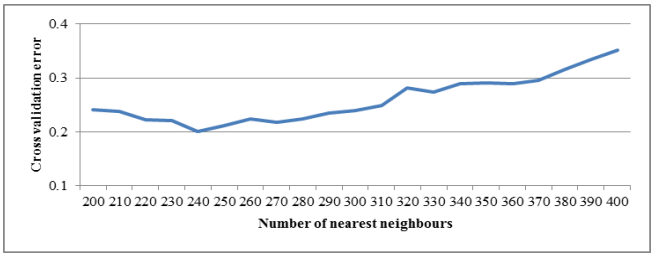


Figure 4. Results of v fold cross validation for road class A

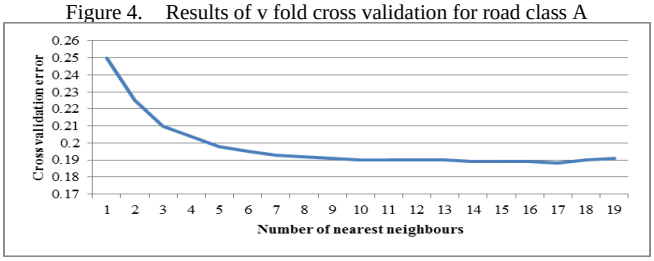


Figure 5. Results of v fold cross validation for road class B

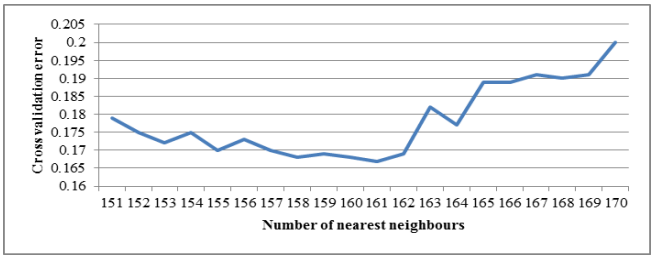


Figure 6. Results of v fold cross validation for road class C

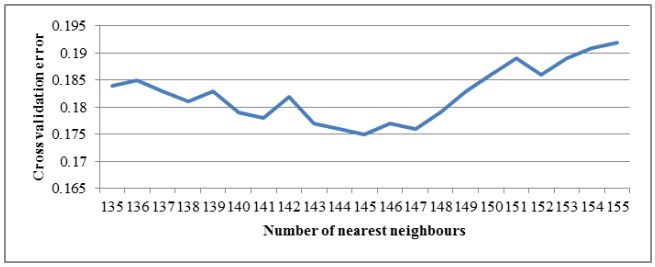


Figure 7. Results of v fold cross validation for road class D

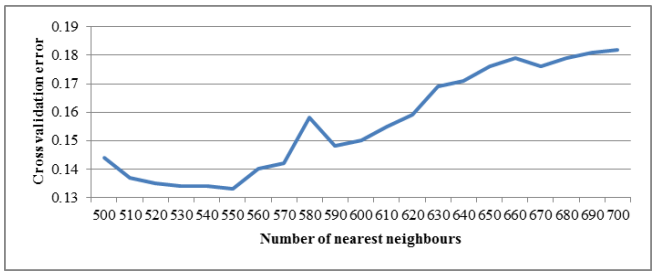


Figure 8. Results of v fold cross validation for road class E

On Figures 4-8, for each road class, computed errors are shown for number of nearest neighbor around optimal value for *k* (optimal in a cross-validation sense). Search for optimal value of *k* was done in phases and figures 4-8 are displaying only results of phase in which optimal *k* was selected for each road class.

D. IRM based approach

Based on previously described procedure, for each road class IRM function is determined and given below:

TABLE 6. IRM FUNCTIONS FOR EACH ROAD CLASS

Class A
$y= 12.85423 + 0.99152*x_1 + 0.09618 * x_2 - 0.0603* x_3 -0.2758*x_4 - 0.0702*x_5 -0.3456*x_6 -0.005*x_7+0.00136*x_8 +0.000074*x_8 +0.02841*x_9-0.0001*x_{10}-0.00004*x_{11}-6.697*x_{12}--0.0404*x_{13}- 0.0453*x_{14}+0.0292*x_{15}-0.0009*x_{16}+0.10499*x_{17}+0.04875*x_{18}$
Class B
$y= 438.20584 + 1.0033*x_1 + 0.00382 * x_2 - 25.58* x_3 -1.442*x_4 - 0.0516*x_5 -0.0519*x_6 -0.0006*x_7+0.63512*x_8 +0.0.12485*x_9 - 0.00258*x_{10}$
Class C
$y= -914.62584 + 1.0033*x_1 + 0.0897 * x_2 - 0.0508* x_3 -0.0216*x_4 +0.21456*x_5 -0.0663*x_6 -0.0498*x_7-0.1783*x_8 -0.04844*x_9 - 0.4439*x_{10} +0.0434*x_{11}$
Class D
$y= 33441.848 + 0.99755*x_1 + 0.01721 * x_2 - 0.0272* x_3 -730.4*x_4 +0.19136*x_5 -0.4659*x_6 -0.0384*x_7$
Class E
$y= 59963.37591 -1405*x_1 + 0.71646* x_2 - 817.8* x_3 -0.55613*x_4 - 0.2512*x_5 -0.00008*x_6 +0.02573*x_7-0.0171*x_8$

Table 6 incorporate variables from all three data sources proving this way their justification as model inputs.

V. URBAN TRAVEL TIME ESTIMATION RESULTS

Based on defined steps, different kNN and IRM regression procedures were applied on data set for each road class. From testing data set a number of vehicle tracks for each time interval is selected and compared with achieved results. Based on this data MAPE (mean absolute percentage error) is calculated for each road class and each time interval defined:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \quad (7)$$

where:

n – number of observations,

A_i – the actual value of recorded travel time for observation i,

F_i – the forecast value of travel time for observation i.

Described results are presented on Figure 9 for morning peak period, Figure 10 for noon time period, Figure 11 for afternoon peak period, and Figure 12 for evening period. For a morning peak period best results are achieved for road class E and IRM method, while for noon time interval best results are achieved for road class A and kNN method. For afternoon time interval best results are achieved for road class D with IRM method, while the same method results in highest MAPE for road class A. Results of evening time interval show that the best results are scored by kNN method for a road class D and the worst ones for road class C by IRM.

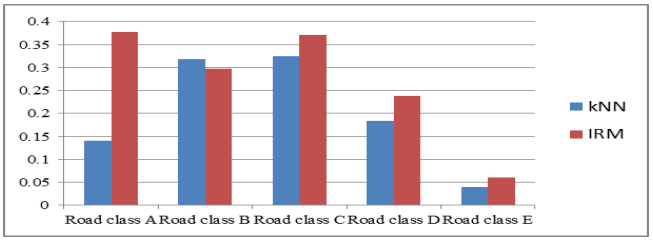


Figure 9. Results of morning peak period

Figure 10. Results of noon period

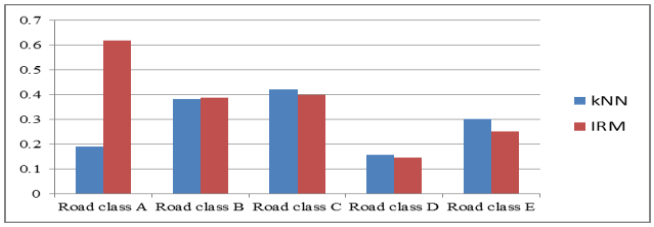


Figure 11. Results of afternoon peak period

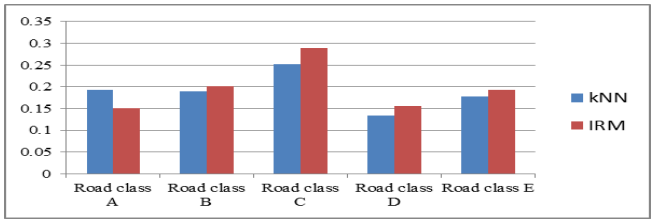


Figure 12. Results of evening period

Method with better results for defined road class and time interval is used for real time travel time estimation and results are presented in Table 7.

TABLE 7. SIMULATION RESULTS FOR EACH ROAD CLASS

Road class	MAPE
Class A	0.043197
Class B	0.188525
Class C	0.089705
Class D	0.052738
Class E	0.141764

VI. CONCLUSIONS

Based on results achieved from case study in City of Zagreb, we can see that GPS tracks, meteorological data and data on road infrastructure present good starting point for urban travel time estimation.

In a case of large data sets application of traditional statistical approach is not advised due to the computationally demanding process of travel time estimation. Dimensionality reduction applied in this paper resulted in reduction of 58 input variables into the average of 21 of them mainly reducing variables with high correlations while keeping maximum data variability contained in original data set. For example, it was evident that data on snow cover thickness and thickness of new snow cover are highly correlated. Data on new snow cover had higher influence on speed deviations and therefore remained in travel time estimation data set after the dimensionality reduction step. Also, data on direction and speed of wind had poor influence on travel time estimation and were mainly removed during dimensionality reduction step. Relative humidity and horizontal visibility were correlated in significant manner together with *y* coordinate of the record during autumn/winter period. This can be explained through influence of river Sava and often fogs in surrounding areas during the cold weather. Since *y* coordinate yield significant correlation with travel time only in this context it was often omitted during dimensionality reduction phase together with horizontal visibility variable while *x* coordinate proved to have high influence on the value of travel time. Beside *x* coordinate, road segment length, wet ground temperature, average speed, speed deviation, log time and daily air temperature remained in data set after dimensionality reduction step for each road class. As expected, values that identified road/segment as road name, segment ID, beginning and end coordinates had very high correlation and were mainly presented by segment ID variable after dimensionality reduction step. Also, it should be noted that vehicle ID remained as significant variable in data set proving the influence of driving characteristics of each driver to the travel time value.

For reduced data sets, kNN and IRM methods results were compared. Based on the achieved results, real time forecasting procedure is applied with forecasting interval of 2 hours. Best results are achieved for road class A (mainly freeways) with MAPE 0.043197, followed with road class D and C. Lowest MAPE is achieved for road class B and it is 0.188525.

Comparison of results to others research is very difficult due to the fact that there is no research that uses same road classification technique or definition of characteristic estimation time interval. Therefore comparison will be made based on most similar values for characteristic time interval and/or road class.

When compared to the results achieved in literature [20] for afternoon peak hour, road length of 2.9 km, desired speed of 64 km/h and without consideration of a long queue that based on this data would correspond to the road class A from our model the error achieved by model proposed in this paper is twice smaller (4.3%) than average error from literature (8.3%).

When results are compared with Google maps [21] travel time estimate for a road class C in length of 2.317 km and afternoon peak period, the achieved result by Google maps is 6 minutes (360 sec) while model proposed in this paper gave estimated travel time for this input data to be 477 seconds. Travel time value measured by GPS record was 476 seconds.

Compared to results for morning peak period MAPE values of three algorithms from literature [22], point-detection-based algorithm (10.6 %), probe-based algorithm with 5% probe rate (10.8%) and adaptive Kalman filter algorithm with 5% probe rate (7.6%), are higher than average MAPE value for all road classes achieved by model proposed in this paper. Average MAPE for all road classes achieved by proposed model equals 5.02% for morning peak period.

ACKNOWLEDGMENT

Authors are grateful to Mireo d.o.o. for GPS vehicle tracks and digital map provision.

Authors are grateful to Meteorological and Hydrological Service for data provision.

 ISBN: 978-1-61208-242-4

REFERENCES

[1] Lum, K.M., Fan, H. S.L., Lam, S.H., and Olszewski, P. (1998) Speed-Flow Modeling of Arterial Roads in Singapore, Journal of Transportation Engineering, Vol. 124, no. 6, pp. 213-222. doi: 10.1061/(ASCE)0733-947X(1998)124:3(213) [retrieved: July, 2012]

[2] J.W.C. van Lint. (2004) Reliable Travel Time Prediction for Freeways. Research School for Transport, Infrastructure and Logistics, Netherlands, 2004.

[3] Richardson, A. J. and Taylor, M.A.P. (1978) Travel time variability on commuter journies, High Speed Ground Transportation Journal, Vol. 12, No.1, pp 77-99.

[4] Chien, S.I-J. and Kuchipudi, C. M. (2003) Dynamic travel time prediction with real-time and historic data, Journal of Transportation Engineering, Vol. 129, No. 6, pp. 608-616., doi: 10.1061/(ASCE)0733-947X(2003)129:6(608) [retrieved: July, 2012]

[5] Cavar, I., Kavran, Z. , and Petrovic, M. (2011) Hybrid approach for urban roads classification based on GPS tracks and road subsegments data, Promet – Traffic&Transportation, 23 (4): pp. 289-296.

[6] Kisgyorgy, L. and Rilett, L.R. (2002) Travel Time Prediction by Advanced Neural Network., Periodica Polytechnica Civil Engineering, Vol. 46, No. 1, pp. 15-32.

[7] Versteegt, H. H. and Tampere, C. M. J. (2005) PredicTime state of the art and functional architecture, TNO Inro report 2003-07, No. 03-7N-024-73196.

[8] Sun, H., Liu, H.X.,Xiao, H. , and Ran,B., (2003) Short term traffic Forecast Using the Local Linear Regression Model. TRB paper no. 03-3580. Transportation Research Board, 82th annual meeting, January 2003., electronic proceedings.

[9] Smith, B.L., Williams, B.M., and Oswald, R.K. (2002) Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research, Part C, Vol. 10 (2002), pp. 303-321. doi: 10.1016/S0968-090X(02)00009-8 [retrieved: July, 2012]

[10] Kamarianakis, Y. and Prastacos, P. (2003) Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches, Transportation research record, vol. 1857, pp. 74-84.

[11] Zhou,C. and Nelson, P.C. (2002) Predicting Traffic Congestion Using Recurrent Neural Networks, 9th World Congress on ITS. Chicago, electronic proceedings.

[12] Tao, Y. and Ran, B. (2006) An Application of Neural Network in Corridor Travel Time Prediction in the Presence of Traffic Incidents, Applications of Advanced Technology in Transportation, 2006., Chicago, Illinois, USA, pp. 455-460, doi 10.1061/40799(213)72) [retrieved: July, 2012]

[13] Vanajakshi, L. and Rilett, L. R. (2007) Travel Time Prediction Using Support Vector Machine Technique, IEEE Intelligent Vehicles Symposium, Istanbul, Turky, pp. 600 - 605.

[14] Wild, D. (1994) Pattern-based Forecasting, Proceedings of the Second DRIVE-II Workshop on Short-Term traffic forecasting, Delft, pp 49-63.

[15] Ben-Akiva, M.E:, Bierlaire, M., Didier, B., Koutsopoulos, H.N., and Rabi, M., (2006) Simulated-based tools for dynamic traffic assignment: DynaMIT and applications, ITS America 10th annual Meeting, Boston, USA, electronic proceedings.

[16] Ben-Akiva, M.E:, Bierlaire, M., Didier, B., Koutsopoulos, H.N., Rabi, M., (2006) DynaMIT: a simulation-based system for traffic prediction, DACCORD Short-term forecasting workshop, Delft, The The Netherland, pp. 19-23.

[17] Webb, A.R.: Statistical Pattern Recognition, John Wiley & Sons, Ltd., 2002.

[18] Ćavar, I., Marković, H., and Gold, H., 2006. GPS vehicles tracks data cleaning methodology, [CD. ROM] Proceedings of the xth ICTS 10th International Conference on Traffic Science.

[19] Spatial plan of City of Zagreb, Prostorni plan grada Zagreba, Službeni glasnik, Zagreb 2009

[20] Liu, H.C., Ma, W., Wu, X., and Hu, H., (2009) Real-Time Estimation of Arterial Travel Time under Congested Conditions, Transportmetrica, Volume 8, Issue 2, 2012., doi: 10.1080/18128600903502298 [retrieved: July, 2012]

[21] http://maps.google.com/ [retrieved: june, 2012]

[22] Chu, L., Oh, J.-S., and Recker, W., Adaptive Kalman filter based freeway travel time estimation, 84th TRB Annual Meeting, Washington D.C., January 9-13 2005. Transportation Research Board. http://www.www.clr-analytics.com/Files/Adaptive%20Kalman%20Filter%20Based%20Freeway%20Travel%20time%20.pdf [retrieved: june, 2012]

 ISBN: 978-1-61208-242-4

Automated Predictive Assessment from Unstructured Student Writing

Norma C. Ming[1, 2]
[1] Nexus Research & Policy Center
San Francisco CA, USA

[2] Graduate School of Education
UC Berkeley, Berkeley CA, USA
Norma@NexusResearchCenter.org

Vivienne L. Ming[3, 4]
[3] Socos LLC
Berkeley CA, USA

[4] Redwood Center for Theoretical Neuroscience
UC Berkeley, Berkeley CA, USA
neuraltheory@socos.me

***Abstract*—We investigated the validity of applying topic modeling to unstructured student writing from online class discussion forums to predict students' final grades. Using only student discussion data from introductory courses in biology and economics, both probabilistic latent semantic analysis (pLSA) and hierarchical latent Dirichlet allocation (hLDA) produced significantly better than chance predictions which improved with additional data collected over the duration of the course. Hierarchical latent Dirichlet allocation yielded superior predictions, suggesting the feasibility of mining student data to derive conceptual hierarchies. Results indicate that topic modeling of student-generated text may offer useful formative assessment information about students' conceptual knowledge.**

Keywords-Predictive assessment; learning analytics; text mining; topic modeling; online discussion.

I. INTRODUCTION

Effective instruction depends on formative assessment to discover and monitor student understanding [1]. By revealing what students already know and what they need to learn, it enables teachers to build on existing knowledge and provide appropriate scaffolding [2]. If such information is both timely and specific, it can serve as valuable feedback to teachers and students and improve achievement [3][4].

Yet incorporating and interpreting ongoing, meaningful assessment into the learning environment remains a challenge for many reasons. Most teachers lack training in assessing understanding beyond the established testing culture [5]. Designed as summative assessments for an outside audience, externally designed tests offer limited information to teachers and students, with reduced opportunities for more varied and frequent assessment, long gaps between taking a test and receiving feedback from it, and often only coarse-grained feedback on the performance of groups of students on broad areas. Even for highly skilled teachers who can infer their students' knowledge from informal assessment activities, aggregating and examining data in detail is both time-consuming and difficult.

Further, testing is often intrusive, demanding that teachers interrupt their regular instruction to administer the test. Assessments that have been developed in conjunction with prepackaged curricula typically require adapting one's own instruction to incorporate at least some of their learning activities. Whether due to differences in state standards, particular student needs, or unique local contexts, teachers may not always be free to adopt externally developed teaching and testing materials.

Our proposed solution to these problems is to build a system which relies on the wealth of unstructured data that students generate from the learning activities their teachers already use. Using automated machine intelligence to analyze large quantities of passively collected data can free up instructors' time to focus on improving their instruction, informed by their own data as well as those of other teachers and students. Building an assessment tool which they can invisibly layer atop their chosen instructional methods affords them both autonomy and information.

This paper describes the design of the system, the data source, and the techniques used. A discussion of the results and their implications for future work follow.

II. DESIGN OF THE SYSTEM

Validating any assessment requires aligning it with outcomes of value. What those outcomes are or should be can vary; our intent here is simply to demonstrate that unstructured student data have predictive value, not to make any claims about what those desirable learning outcomes should be. As a proof of concept, we are predicting end-of-course grades, although the same approach may be applied to many other assessments.

As inputs, our system relies on what students actually do, rather than information associated with their identities and backgrounds. Other predictive analytics systems include data on demographics, schooling history, and measures of motivation [6][7], variables which have been shown to predict student retention and performance but which also may reflect prior social, economic, or cultural inequities. Since our goal is to provide predictive information to the teachers and students, we hope to avoid exacerbating these inequities by minimizing the visibility (but not denying the reality) of these influences [8].

Numerous other academic analytics systems incorporate measures of student activity and course performance [9][10] [11][12]. We seek to go beyond simple quantity-based metrics of effort, participation, and engagement, by analyzing the semantic content of student-generated products in order to assess what students know, not how active they

 ISBN: 978-1-61208-242-4

are. This also enables deeper insights into the ideas which students are addressing, rather than vocabulary, punctuation, sentence complexity, or other linguistic features that signal writing quality [13]. While many computer-aided or intelligent tutoring systems incorporate sophisticated analyses of students' performance on prespecified problems [14], our goal is to explore nuances in student knowledge from a wider diversity of learning experiences.

To elucidate the semantic content of unstructured text data, we employ probabilistic latent semantic analysis (pLSA) and hierarchical latent Dirichlet allocation (hLDA) [15][16]. These techniques yield topic models of the student-generated texts by analyzing word co-occurrence within documents, specifically discussion forum posts in this case. Both pLSA and hLDA are generative, probabilistic models which provide low-dimensional descriptions of text by inferring small sets of latent factors, or topics, which explain the distribution of words in the analyzed documents. Both employ the simplifying "bag-of-words" assumption that a document can be represented as an unordered count of words. Each document is "generated" by mixing topics and then selecting words from those topic mixtures. We expanded on this by using collocation information to automatically select a set of domain-specific phrases (*n*-grams) of arbitrary word length [17]. Topics then describe the distribution of these *n*-grams, including single words and phrases.

For pLSA, the distributions describing topic and *n*-gram likelihoods are assumed to be Gaussian, providing a simple model of document composition. Although pLSA has its limitations, it and its predecessor LSA are widely used to model the semantic content of text. We will use the topics it infers from student posts, specifically the inferred coefficients of the latent factors, as the predictor variables for students' final grades. In other words, over the duration of a multi-week class, do the concepts discussed by students as inferred by pLSA predict their course outcomes? If so, how does the accuracy of these predictions change over time as more student work is analyzed?

The hLDA model provides a much more complex model of the text, combining a more reasonable multinomial model of word occurrence with the ability to infer an arbitrary hierarchy to the topics. Documents are not just mixtures of topics but mixtures of topic branches in a semantic tree structure. A draw from the distribution over branches defines a general topic, while a draw from the distribution of branch depth defines the specificity of the *n*-gram within a branch, such as science → biology → neuroscience → sensory neuroscience → cortical vision → etc. This hierarchical organization, as learned from the data, provides an additional piece of information to test for our outcome predictions. That is, does the specificity of a topic in student posts, as represented by depth in the hLDA tree, aid in predicting course outcomes?

TABLE I. SUMMARY STATISTICS ON THE TWO DATASETS ANALYZED

	Biology	**Economics**
Course length (in weeks)	5	6
# of discussion question threads per class	10	12
# of classes	17	45
# of students (after filtering)	230	970
# of posts by students	9118	44345

III. DATASET AND METHODS

We independently applied pLSA and hLDA to archived data from the online discussion forums of two introductory courses at a large for-profit university, an undergraduate biology course and an MBA-level economics course (**Table 1**). The biology course sample was limited to focus on instructors analyzed in previous research [18], while the economics course sample represented all such classes available in a particular archived database. Throughout both courses, students were required to respond to two discussion questions per week in these forums. Although individual instances (classes) of a course could vary slightly in the specific questions and assignments posed to students, all adhered to a standard course outline and schedule with regard to learning goals, topics covered, and texts used. We removed data from students who dropped out before earning a final grade and normalized final grades to be between [0,1].

We trained a logistic regression model to predict the final grades based on the accumulated weekly topic coefficients. While logistic regression certainly will not yield best-in-class performance, its simplicity and transparency allow for a cleaner analysis of the topic models. Both models were trained in batches with the student posts using the same *n*-gram dictionaries. No normative material was used for training, only student posts. We used five-fold cross-validation and trained pLSA and hLDA on individual posts independently for each course. The data were partitioned into five sets, and on five separate rounds of training and test, a different set was held out for testing while the remainder was used for both unsupervised and supervised training. The results reported below were averaged across the five training rounds.

The results from pLSA had 30 factors, while hLDA produced a tree composed of 50 nodes with a maximum depth of four layers. Once the pLSA topics and hLDA trees were learned, we took weekly posts by the students and projected them into topic space and the semantic tree, respectively. We then used the inferred coefficients for the topic factors from the current week's posts, concatenated with any from previous weeks, as predictor variables. For hLDA, each topic produced pairs of inferred coefficients representing both the loading of the topic and the topic's depth in the tree hierarchy.

 ISBN: 978-1-61208-242-4

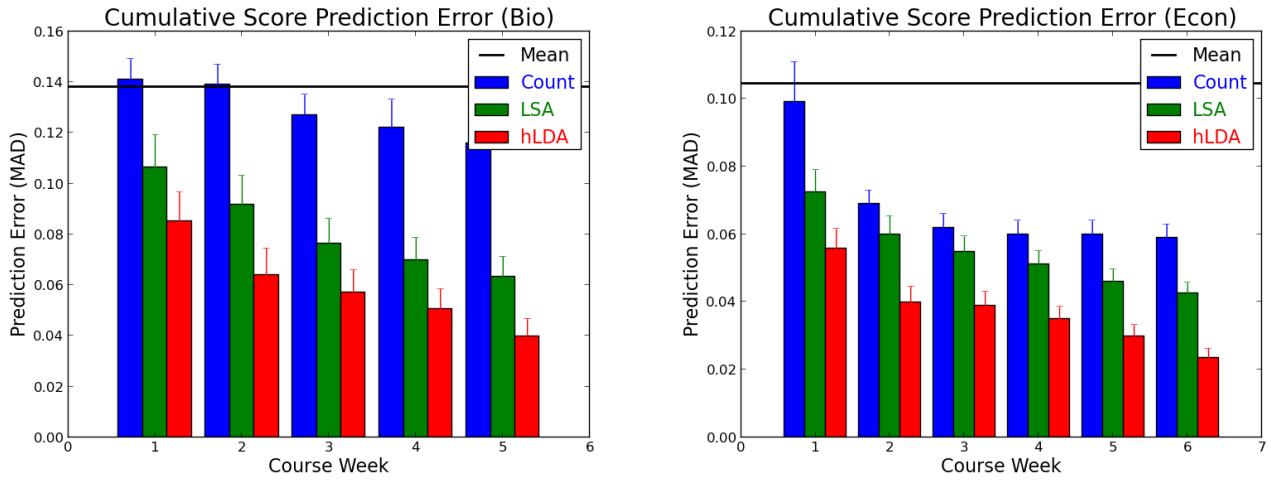


Figure 1. Accuracy of pLSA and hLDA in predicting students' final grades from the topics in their discussion posts (MAD = mean absolute deviation).

IV. RESULTS

The results shown in Figure 1 illuminate all three of our research questions regarding the efficacy of topic modeling in predicting students' grades. The graphs depict the mean absolute deviation (MAD) between students' predicted and actual final grades for each model over the duration of the course. For reference, the black line shows the error that would result from predicting the course mean, and the blue bars show the prediction based on word count per post.

The first finding is that topic modeling using pLSA produces significantly better than chance predictions of students' course grades, even from the first week of the course. Topic modeling also produces consistently better predictions than post length for the biology data, with a smaller advantage for the economics data. Second, accumulating data over additional weeks of the course yields significant modest gains. For example, by the end of the economics course, the pLSA model's prediction is approximately within ±0.05 of the actual grade (or within one letter grade). Third, the hierarchical modeling of hLDA gives better predictions than pLSA.

Additional examination of the data also reveals that higher course grades are correlated with a slightly higher mean of the depth parameter in hLDA. Topics in the hLDA model are structured in a hierarchy learned from the data, with more specialized topics being represented deeper in the hierarchy than more general topics. The central topic is the most general language that appears in all documents, and thus appears at the topmost level (lowest level of depth) in the hierarchy. Figure 2 depicts the percentage of *n*-grams used by students receiving letter grades of A, B, and C at each of the four depth levels specified in the hierarchy. As shown, most of the language used by students who receive C's resides at the topmost level, while relatively greater percentages of the language used by students receiving A's and B's reside at deeper levels in the hierarchy.

A preliminary reading of selected discussion posts indicated that higher grades correlated with more technically proficient language use. Posts containing more general language tended to include more anecdotal comments, whereas posts with more technically specific language addressed course concepts in greater depth. Deeper analysis of potential relationships between these metrics and post quality will be valuable for elucidating how hierarchy depth may correspond with discussion and course characteristics of interest.

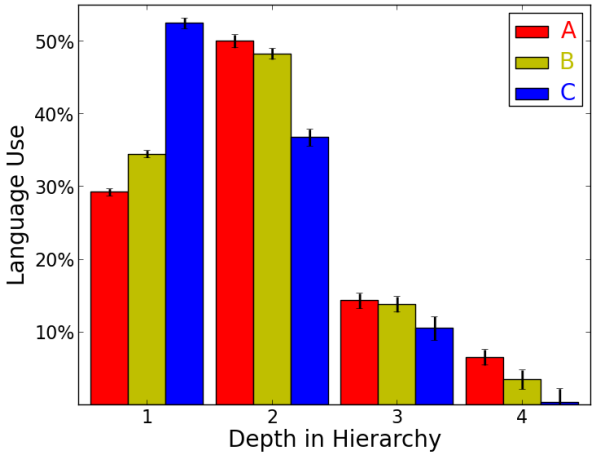


Figure 2. Correlation between language depth and course grade.

V. CONCLUSION AND FUTURE WORK

This work demonstrated that unstructured student data in the form of discussion forum posts can be used to predict assessment outcomes of interest (final course grades). It extends previous research investigating LSA and related computational approaches to predicting student outcomes from text data [19][20][21], illustrating the predictive value

 ISBN: 978-1-61208-242-4

of adding more data over time as well as the utility of hierarchical topic modeling. The improvements over the duration of the course may reflect the benefits of including a broader range of course concepts and using more recent student performance data. The advantages of hLDA reveal the possibility and benefit of algorithmically discovering domain-specific conceptual hierarchies in student-generated text. The correspondence between hierarchical depth and course grade suggests just one dimension of knowledge which hierarchical modeling reveals from students' writing; further analysis may enable identifying and interpreting other dimensions.

As such these methods show potential for application as a type of formative assessment, to provide more content-relevant feedback to students and teachers about students' thinking in order to better guide learning and instruction. Continued research will be worthwhile for exploring the impact of changes to the algorithms, as well as the inputs (*e.g.*, essays, responses to short-answer questions) and outcomes (*e.g.*, course retention, scores on exams or other assignments). Additional steps include exploring how to present this feedback usefully and possible interventions for students and teachers to follow. While we opted to focus on semantic content alone for this project, future work may investigate the relative value of combining semantic data with other features of performance and additional student data.

REFERENCES

[1] J. W. Pellegrino, N. Chudowsky, and R. Glaser, Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academy Press, 2001.

[2] L. A. Shepard, The role of classroom assessment in teaching and learning. (CSE Technical Report 517). Los Angeles CA: Center for the Study of Evaluation, 2000.

[3] A. N. Kluger and A. DeNisi, "Effects of feedback intervention on performance," Psychological Bulletin, vol. 119(2), 1996, pp. 254-284.

[4] P. Black and D. Wiliam, Assessment and classroom learning, in Assessment in Education: Principles, Policy, and Practice, vol. 5(1), 1998, pp. 7-74.

[5] M. C. Ellwein and M. E. Graue, "Assessment as a way of knowing children," in Making schooling multicultural: Campus and classroom, C. A. Grant and M. L. Gomez, Eds. Englewood Cliffs, NJ: Merrill, 1996.

[6] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," EDUCAUSE Review, vol. 42(4), 2007, pp. 41-57.

[7] L. V. Morris, S. S. Wu, and C. Finnegan, "Predicting retention in online general education courses," American Journal of Distance Education, vol. 19(1), 2005, pp. 23-36.

[8] C. M. Steele and J. Aronson, "Stereotype threat and the intellectual test-performance of African-Americans," Journal of Personality and Social Psychology, vol. 69(5), 1995, pp. 797-811.

[9] K. E. Arnold, "*Signals*: Applying academic analytics," EDUCAUSE Quarterly, vol. 33(1), 2010.

[10] E. J. M. Lauría and J. Baron, Mining Sakai to measure student performance: Opportunities and challenges in academic analytics, 2011. Retrieved from http://ecc.marist.edu/conf2011/materials/LauriaECC2011-%20Mining%20Sakai%20to%20Measure%20Student%20Performance%20-%20final.pdf

[11] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an ''early warning system" for educators: A proof of concept," Computers and Education, vol. 54, 2010, pp. 588-599.

[12] H. Zhang, K. Almeroth, A. Knight, M. Bulger, and R. Mayer, "Moodog: Tracking Students' Online Learning Activities," in Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Vancouver, Canada, 2007.

[13] S. A. Crossley, R. Roscoe, A. Graesser, and D. S. McNamara, "Predicting human scores of essay quality using computational indices of linguistic and textual features," in G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, New Zealand: AIED, 2011, pp. 438-440.

[14] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, vol. 46(4), 2011, pp. 197-221.

[15] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, pp. 50-57.

[16] D. Blei, T. Griffiths, and M. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," Journal of the Association for Computing Machinery, vol. 57(2), 2010, pp. 1-30.

[17] D. Blei and J. Lafferty, Visualizing topics with multiword expressions, 2009. arXiv:0907.1013

[18] N. C. Ming and E. P. S. Baumer, "Using text mining to characterize online discussion facilitation," Journal of Asynchronous Learning Networks, vol. 15(2), 2011.

[19] B. Rehder, M. E. Schreiner, M. B. W. Wolfe, D. Laham, T. K. Landauer, and W. Kintsch, "Using latent semantic analysis to assess knowledge: Some technical considerations," Discourse Processes, vol. 25(2-3), 1998, pp. 337-354.

[20] P. W. Foltz, D. Laham, and T. K. Landauer, "Automated essay scoring: Applications to educational technology," in Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, B. Collis and R. Oliver, Eds. Chesapeake VA: AACE, 1999, pp. 939-944.

[21] M. J. Ventura, D. R. Franchescetti, P. Pennumatsa, A. C. Graesser, G. T. Jackson, X. Hu, Z. Cai, and the Tutoring Research Group. "Combining computational models of short essay grading for conceptual physics problems," in Intelligent Tutoring Systems, J. C. Lester, R. M. Vicari, and F. Paraguacu, Eds. Berlin, Germany: Springer, 2004, pp. 423-431.

 ISBN: 978-1-61208-242-4

Building OLAP Data Analytics by Storing Path-Enumeration Keys into Sorted Sets of Key-Value Store Databases

Luis Loyola
R&D Division
SkillUpJapan Corporation
Tokyo, Japan
loyola@skillupjapan.co.jp

Fernando Wong
R&D Division
SkillUpJapan Corporation
Tokyo, Japan
f.wong@skillupjapan.co.jp

Daniel Pereira
R&D Division
SkillUpJapan Corporation
Tokyo, Japan
d.pereira@skillupjapan.co.jp

***Abstract*—In this paper, a novel approach that allows the retrieval of OLAP analytics by means of storing multidimensional data in key-value databases is described and evaluated. The proposed mechanism relies on generated keys based on the path-enumeration model built upon a tree representation of a relational database. Each key represents a combination of OLAP dimension labels and is inserted into a key-value sorted set where its associated values are the metrics of interest, i.e., the OLAP *facts*. Our implementation for a real advertisement server system in Redis and its performance comparison with a relational OLAP (ROLAP) database running on top of MySQL, show that our proposed scheme can answer complex multidimensional queries much faster, while keeping a great flexibility of the data schema. In contrast with general ROLAP schemes, which usually require the pre-computation of tables for specific queries, the sorted sets in the proposed system are not pre-computed but generated on demand, so no further delay is introduced. To the best of the authors' knowledge, this is the first system that deals with mapping relational data structures into sorted sets of key-value store databases to answer complex queries on multidimensional data.**

Keywords-OLAP; path-enumeration keys; sorted sets; databases; key-value stores.

I. INTRODUCTION

By making use of object-relational mapping (ORM) [9], object-oriented programming classes on the application level can be represented in a persistent way on a relational database. With ORM, the Entity Relation map of a relational database can be represented by a parent-child relationship tree. In this type of trees, relations between nodes can be represented using verbal expressions like *"has one"*, *"has many"*, *"has and belongs to many"*, *"belongs to"* and so on. In fact, these expressions describing relationships among classes are very popular in abstraction layers of SQL data connectors developed for mainstream scripting languages like Java[2], Python[5] or Ruby[8].

Online Analytical Processing (OLAP) databases are a key tool for the analysis of large amounts of data. An OLAP [20] database corresponds to a multidimensional database where measures or statistics of interest are pre-calculated in the so-called *fact tables*. The flexibility and high-speed with which *multi-dimensional OLAP* (MOLAP) allows to perform complex queries involving several dimensions have turned it into one of the most useful tools for modern data analysis. On the other hand, during the last decade there has been a big hype on no-SQL databases, featuring various software solutions and open-source tools. The technologies, concepts and algorithms used in them differ quite substantially but one can observe the presence of the key-value store concept in many. The basics of such key-value store systems are databases where a hash key can be associated to a value, which can range from a simple text string to more complex data structures such as arrays, lists or even binary files.

This paper introduces an algorithmic approach that maps a relational database into tailored sorted sets in a key-value store database. In a cold start, no data needs to be transferred from the relational database to the proposed system since from the application's point of view it works as a caching layer, not needing to pre-compute tables as they are generated on the fly as the queries arrive. To the best of the authors' knowledge, this is the first system that deals with automatically mapping relational data structures into sorted sets of key-value store databases to answer complex queries on multidimensional data like in OLAP.

The proposed data representation structure is helpful for analyzing multidimensional data in real-time as in OLAP. The advantage of the proposed approach is two-folded: 1) on the practical side, OLAP solutions are quite costly and complex, so they usually require a large investment on dedicated servers, software licenses and staff training. In contrast, there exist a handful of excellent open-source enterprise-level key-value store databases which are easy to configure and use, and those are the target tools of this work. 2) On the technical side, the mechanism proposed in this document differs substantially to the OLAP approach for consolidating the multidimensional data structure for analytics, and furthermore allows the user to still keep her relational database in place to store the almost-static data.

The paper is organized as follows: in Section II, we review some key concepts. Works related to our approach are briefly mentioned in Section III. Section IV provides a more in-depth description of the proposed mechanism. Section V

 ISBN: 978-1-61208-242-4

shows performance measurements and comparison with a ROLAP database running on top of MySQL in a real implementation. Finally, some final remarks are provided in Section VI.

II. BACKGROUND

In this section, a brief overview of key concepts like OLAP, sorted sets, path enumeration as well as a brief description of the Redis database are provided.

A. OLAP

OLAP stands for Online Analytical Processing, a category of software tools that provides analysis of data stored in a database. OLAP tools enable users to analyze different dimensions of multidimensional data, and allow flexible and fast access to the data and those multiple dimensions. Such dimensions are stored as a side of what is called an OLAP cube, or hypercube, building that way a multidimensional cube which stores information, also known as facts or measures. Such facts store no other than the aggregated totals of relevant key information that one wants to analyze. All this data is, typically, created from a star or snowflake schema of tables in a relational database (RDB) such as the simplified one shown in Figure 1, which depicts the RDB for a real advertising server application deployed by our company. Our star schema features six *dimensions*, represented as the peripheral tables, and we are interested in measuring the aggregated results from the central *facts* table, all relevant to the presented dimensions. Each *measure* can be thought of as having a set of *labels*, or meta-data associated with it. A *dimension* is what describes these *labels*; it provides information about the *measure*. As a practical example, in our systems, a cube contains the clicked advertisements as a measure and the advertisement space as a dimension. Each clicked advertisement can then be correlated with its campaign, publisher, advertiser, time, campaign resource dimensions and any other number of dimensions can be added, as long as data to correlate those dimensions with the click are added in the structure, such as a foreign key for instance. This allows an analyst to view the *measures* along any combination of the *dimensions*. Additionally, OLAP systems are typically categorized under three main variations:

- MOLAP
  It is the standard way of storing data in OLAP and that is why it is sometimes referred to as just OLAP. It stores the data on a multi-dimensional array storage where fact tables are pre-computed. It does not rely on relational databases.
- ROLAP
  It works on the top of relational databases. The base data and dimension tables are stored as relational tables and new tables are created to store the aggregated information. It gives the appearance of traditional slicing and dicing OLAP functionalities working with the data stored in the relational database.
- Hybrid
  There is no clear definition of what a "hybrid" OLAP system means, but in all cases the data is stored in both a relational database and special array storage with pre-computed data.

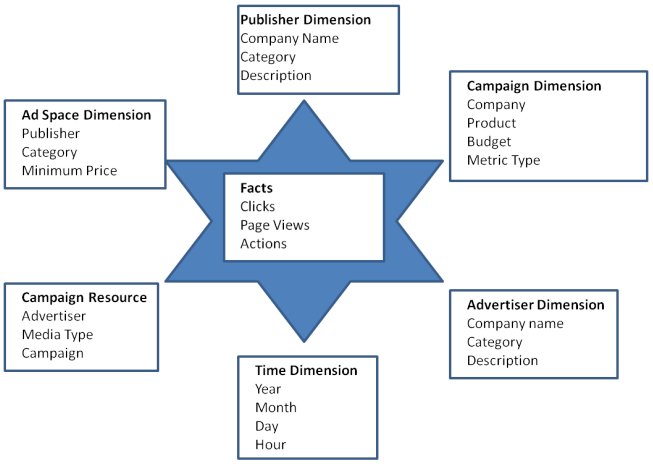


Figure 1. Sample star data model in OLAP.

B. Sorted Sets

A sorted set is a data structure defined by a list where each element has an associated score. In contrast with a normal set, a sorted set can be ordered based on its elements' score. Hence, a sorted set $\mathbf{Z} = \{(e_1, s_1), \ldots, (e_n, s_n)\}$, where each element e_i has score s_i, can be represented as a hash where each key has an associated score value and the elements in the set can be ordered by their score.

C. Path Enumeration Model

One of the properties of trees is that there is one and only one path from the root to every node. The path enumeration model stores that path as a string by concatenating either the edges or the keys of the nodes in the path. Searches are done with string functions and predicates on those path strings. There are two methods for enumerating the paths, edge enumeration and node enumeration. Node enumeration is the most common, and there is little difference in the basic string operations on either model. In this paper we apply the node enumeration method.

D. Redis

Redis [7] is an open-source memory-only key-value store database with a handful of useful data structures. Similar to other key-value data stores, Redis has as data model a dictionary where keys are mapped to values, with the essential differences that values can be other than strings of text and that it can handle collections, or sets, in an

 ISBN: 978-1-61208-242-4

unordered and ordered manner, which behave in the same way as the sorted sets presented in Section II-B.

At this moment, there is no effective way to distribute Redis other than manually distributing chunks of data in independent servers, typically known as *sharding*. Therefore, measuring the scalability of our proposed scheme highly depends on future work inherent to Redis development. As Redis stores the whole database in memory it is faster than conventional solutions, while maintaining some control over the data persistence. If needed, Redis allows us to take *snapshots* of data as well as asynchronously and periodically transfer data from memory to disk, however this is not relevant to the idea presented in this work. Finally, another benefit of using Redis is the lack of necessity to perform any data schema alteration providing us with higher flexibility to introduce changes on demand.

III. Related Work

Related work for OLAP analytics in key-value store databases is primarily focused on distributed algorithms like MapReduce, cloud services and in-memory OLAP architectures. Bonnet et al. [14], Qin et al. [28], Wang et al. [29], analyze and improve upon existing MapReduce algorithms while Abelló et al. [12] make use of distributed data stores like Bigtable [18], GFS [23] or Cassandra [26]. The reasoning behind the adoption of those architectures is due to the large amount of resources needed when dealing with large data sets. Such architectures also allow parallelization of queries and provide easy means of hardware scaling. However, solutions based on MapReduce are less efficient, due to lack of efficient indices, proper storage mechanisms, lack of compression techniques and sophisticated parallel algorithms for querying large amounts of data [27].

Some related literature focuses on using cloud services, like Dynamo [22] from Amazon, to improve upon current resource utilization and reducing costs [15], [16], [19], as several cloud platforms began to give access to cheap key-value stores that allow easy scaling. Such cloud services can, however, be problematic as privacy issues can arise, if the data cannot be anonymized [10]; data is often stored in untrusted hosts and replicated across different geographic zones, without much control from the users. Moreover, those cloud services lack important data structures, such as the presented sorted sets, for instance. Also, they are proprietary and closed to improvements from the developers, making it impossible to add or develop features when needed or customize a certain part of the engine.

More literature [17], [21] emphasizes the need to combine the benefits of NoSQL database systems with traditional RDBMS, in order to effectively handle big data analytics. With the growth of data stored in RDBMS databases, it is deemed impossible to cover all the needed scenarios and still expect timely reporting of results. By using key-value store databases, data analytics can be offloaded from RDBMS, specially if the operations are simple lookups of objects, with no urgency of having consistent data at a very granular level. Recently, the decline in memory price has allowed for new solutions, entirely based on memory, to become financially feasible. Therefore, architectures providing in-memory OLAP capabilities have arisen, such as DRUID [1], QlikTek [6], SwissBox [13] architecture, H-Store [24] and Cloudy [25]. However, not much detail on their architectures and internal implementation aspects are known and no comparison to any state of the art solution has been published. Most of the solutions provide very controlled environments, often specific to an operating system and are business oriented.

IV. Proposed Scheme

In a tree representing parent-child-relation classes mapped into an underlying relational database, we can generate a basic unique key for each branch by hierarchically listing in a top-to-bottom order the node ids found all the way down from the root to the bottom leafs along every branch. This is equivalent to picking up the deepest paths of every branch from the path enumeration model applied to general trees. In order to store all possible combinations of relevant dimensions for our reports it is essential that every path from the path enumeration model has an associated key. This defines a set of deepest paths $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}$ for n root-through-bottom-leaf paths where each $\mathbf{d}_i$ corresponds to a path represented by a list of nodes ordered from top to bottom in the tree hierarchy. Each of the nodes in the tree represents a dimension that can be included in the reporting system. After the set of deepest paths **D** is constructed, another set $\mathbf{D'} \supseteq \mathbf{D}$ is defined, so that $\mathbf{D'} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n, \mathbf{d}_1\mathbf{:}\mathbf{d}_2, \mathbf{d}_1\mathbf{:}\mathbf{d}_3, \ldots, \mathbf{d}_1\mathbf{:}\mathbf{d}_2\mathbf{:}\mathbf{d}_3\mathbf{:}\ldots\mathbf{:}\mathbf{d}_n\}$ and $|\mathbf{D'}| \leq (2^n - 1)$. Eventually, **D'** should only hold the combination of paths that are frequently consulted via the reporting system.

```
sorted_sets = {}
foreach path ∈ D' do
  | Insert path(time_slot) into sorted_sets
end
```

Algorithm 1: Initialization of sorted sets.

In the next step, one has to concentrate on the statistical values or parameters that should be measured, i.e., the so-called facts in OLAP. After identifying the facts we define sorted sets combining the time dimension (in the maximum resolution of interest) with the paths of **D'**. Thus, the name of each sorted set can be built as specified on Algorithm 1 where *time_slot* represents one time slot in the maximum level of granularity for the reporting system. Every time there is an event related to one of the facts, the corresponding fact's score is incremented for the combination of all dimension ids related to the event. Hence, when

 ISBN: 978-1-61208-242-4

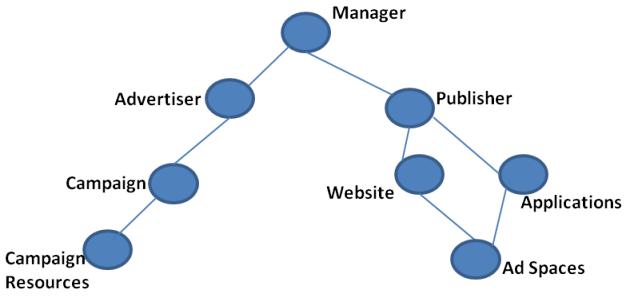


Figure 2. Example entity relationship diagram.

an event measured by a fact occurs, the score of a fact related to a particular combination of dimension ids, denoted as **dimension_ids**$(path)_{fact}$, in sorted set **path**$(time\_slot)$ is incremented by one in case the fact corresponds to an aggregate statistic. If the fact is not measured by an aggregate statistic, the score is totally ignored for all set operations so the sorted set behaves just as a normal set.

For the sake of clarity, an example of the proposed data structure applied to a real advertisement server application is presented. Figure 2 shows the example data structure where the root node "Manager" corresponds to one advertising agency administrator. Each advertising agency has "Publishers" and each of those publishers has "Websites" and smartphone "Applications". In turn, each website and each smartphone application has one or many "Ad Spaces" which are sold to "Advertisers". On the other hand, each advertiser has "Campaigns" of products it wants to promote. Inside each campaign there are many advertising media assets like image banners, video banners, text banners, etc., which are called "Campaign Resources".

In the tree of the example shown in Figure 2, there exist three deepest paths given by the set **D** = { **M:A:C:R**[1], **M:P:W:S**[2], **M:P:Ap:S**[3]}. This set has to be united with the deepest path keys' combination set **E**, which corresponds to **E** = { **M:A:C:R:P:W:S**[4], **M:A:C:R:P:Ap:S**[5] } in our example. After the union operation has been performed, the resulting set contains five unique basic keys that can be derived in a straightforward manner from both sets above as:

$$\mathbf{D'} = \mathbf{D} \cup \mathbf{E} = \left\{ \begin{array}{c} \mathbf{M:A:C:R}, \mathbf{M:P:W:S}, \\ \mathbf{M:P:Ap:S}, \mathbf{M:A:C:R:P:W:S}, \\ \mathbf{M:A:C:R:P:Ap:S} \end{array} \right\} \quad (1)$$

In order to construct the full keys, we need to combine these five basic keys with each of the parameters or statistical measures we are interested in assessing, i.e. the facts in the OLAP jargon as well as the time expressed in the maximum resolution required by our reporting system. Let us say that in the example of Figure 2 we are interested in measuring the views and clicks as fact events. Thus, in the same way as in OLAP, our reporting system should be able to answer all types of multidimensional queries like: "what was the most clicked campaign resource from Advertiser **a** on ad spaces from Publisher **p** between 2011/07/19 and 2011/09/21?", or "what was the number of accumulated clicks and views on Campaigns from Advertiser **a** for all the Ad Spaces from Publisher **p**'s Web Sites between 2011/06/15 and 2011/12/21?", "what was the total number of accumulated clicks and views on campaign resource **r** when shown in Ad spaces from all applications of Publisher **p**?", etc. To build such a reporting system we make use of sorted sets. In order to name the sorted sets we need to follow the steps described below. For all deep-most branches in the tree and their combinations we name their associated sorted sets using the pattern **basic_key**(time_slot). If the time slot corresponds to one hour (i.e. the maximum resolution for accumulated statistics we are willing to tolerate is one hour) the sorted sets associated to the time period between 16:00 and 17:00 of 2011/07/09 can be named as **basic_key**(2011-07-09-16). Inside this sorted set we will increment the score of a particular combination of dimension ids based on the event-related fact. The pseudo-code below describes the basic logic to increment the score of the related basic-key elements in the sorted set.

```
if event = "view" then
    for basic_key ∈ D' do
        [dimension_ids(basic_key)_views].score++
    end
else if event = "click" then
    for basic_key ∈ BasicKeySet do
        [dimension_ids(basic_key)_clicks].score++
    end
end
```

Algorithm 2: Incrementing event-related facts.

As can be seen in Algorithm 2, after naming the sorted sets, we increment by one the score for the combination of ids associated with the basic_key and the related event. For example, let us suppose there is a click at 4:45 p.m. on Campaign Resource **r** from campaign **c** ran by Advertiser **a** (belongs to manager **m**) when shown at an ad space **s** from website **w** of publisher **p**. Algorithm 3 shows how each of the five **basic_key** ∈ **D'** would be modified.

The above sets are suitable for the most granular questions over multiple dimensions for our reporting system. But what happens if we need to know, for example, the total accumulated clicks on website **w** of Publisher **p** during the

[1]Manager, Advertiser, Campaign, Campaign Resource
[2]Manager, Publisher, Web Sites, Ad Spaces
[3]Manager, Publisher, Applications, Ad Spaces
[4]Manager, Advertiser, Campaign, Campaign Resource, Publisher, Web Site, Ad Space
[5]Manager, Advertiser, Campaign, Campaign Resource, Publisher, Application, Ad Space

 ISBN: 978-1-61208-242-4

$TimeSlot = \{$2011-07-09-16, 2011-07-09, 2011-07, 2011$\}$
foreach *t in TimeSlot* **do**
  In set **M:A:C:R**(t)
    [**m:a:c:r**$_{clicks}$].score++
  In set **M:P:W:S**(t)
    [**m:p:w:s**$_{clicks}$].score++
  In set **M:P:Ap:S**(t)
    nothing is modified
  In set **M:A:C:R:P:W:S**(t)
    [**m:a:c:r:p:w:s**$_{clicks}$.].score++
  In set **M:A:C:R:P:Ap:S**(t)
    nothing is modified
end

Algorithm 3: Modification of scores in basic key sets.

month of February 2011. In that case, we can obtain the requested measure by performing union operations among the sorted sets between $date_1 =$ 2011-02-01-00 and $date_2 =$ 2011-02-28-23. In this paper, we present three different approaches to deal with this type of queries. By default, we assume the scores of the same element are added during a union operation of sorted sets, however, the mathematical operation performed on the values during a union can be extended to more complex operations.

A. Approach 1

This approach consists of performing union operations only on the deep-most branch keys of the basic key set **D'**. For our example, the set

$$\mathbf{Q} = \bigcup_{d=date_1}^{date_2} \mathbf{M{:}P{:}W{:}S}(d) \tag{2}$$

contains the total number of accumulated clicks for all ad spaces throughout all days within the specified period for all combinations of publishers, websites and ad spaces. Thus, in order to answer the query stated above, we just have to add the scores of values **m:p:w:***$_{clicks} \in$ **Q**, where $*$ symbolizes all ad spaces that belong to web site **w** of publisher **p** with manager **m**.

The disadvantage of the method above is that we need to search for all elements with pattern **m:p:w:***$_{clicks} \in$ **Q** using regular expressions or other parsing techniques, and then add all their associated scores. This is time-consuming and expensive in terms of resources as we need to load all elements in **Q** to memory and perform the search operation for the pattern in a separate application. In order to address this problem, we store a **M:P:W:S**$(\mathbf{m}, \mathbf{p}, \mathbf{w}, *)$ set, which contains all combinations of the requested dimension ids in the form of **m:p:w:***$_{clicks}$. Then, we can simply perform an intersection between **M:P:W:S**$(\mathbf{m}, \mathbf{p}, \mathbf{w}, *)$ and **Q**, and sum the scores of the members in the resulting set

$$\mathbf{Q'} = \mathbf{M{:}P{:}W{:}S}(\mathbf{m}, \mathbf{p}, \mathbf{w}, *) \cap \mathbf{Q} \tag{3}$$

In practice, we map the function **M:P:W:S(m,p,w,*)** to a set in Redis whose key follows the pattern **M:m:P:p:W:w:S**. Please note that **M:P:W:S(m,p,w,*)** is a regular set, while **Q** and **Q'** are sorted sets. In our context, the intersection between a sorted set and a regular set results in a sorted set whose members are present in both sets, while retaining the score they had on the sorted set. For example, let us assume that the clicks registered between $date_1$ and $date_2$ are contained in the sorted set

$$\mathbf{Q} = \left\{ \begin{matrix} \mathbf{m_1{:}p_1{:}w_1{:}s_1}_{clicks}{=}2, \mathbf{m_1{:}p_1{:}w_1{:}s_2}_{clicks}{=}7, \\ \mathbf{m_1{:}p_2{:}w_2{:}s_3}_{clicks}{=}4, \mathbf{m_1{:}p_2{:}w_2{:}s_5}_{clicks}{=}5, \\ \mathbf{m_2{:}p_3{:}w_3{:}s_7}_{clicks}{=}3, \mathbf{m_2{:}p_3{:}w_4{:}s_8}_{clicks}{=}6 \end{matrix} \right\} \tag{4}$$

and that all combinations of ids for web site $\mathbf{w}_2$ of publisher $\mathbf{p}_2$ with manager $\mathbf{m}_1$ that have so far experienced events such as clicks and views during the entire history of the system are given by

$$\mathbf{M{:}P{:}W{:}S}(\mathbf{m}_1, \mathbf{p}_2, \mathbf{w}_2, *) = \left\{ \begin{matrix} \mathbf{m_1{:}p_2{:}w_2{:}s_3}_{clicks}, \\ \mathbf{m_1{:}p_2{:}w_2{:}s_4}_{clicks}, \\ \mathbf{m_1{:}p_2{:}w_2{:}s_5}_{clicks}, \\ \mathbf{m_1{:}p_2{:}w_2{:}s_6}_{clicks} \end{matrix} \right\} \tag{5}$$

Then, the resulting set **Q'** is denoted by

$$\mathbf{Q'} = \left\{ \mathbf{m_1{:}p_2{:}w_2{:}s_3}_{clicks}{=}4, \mathbf{m_1{:}p_2{:}w_2{:}s_5}_{clicks}{=}5 \right\} \tag{6}$$

The amount of clicks for this particular combination of manager, publisher and website between $date_1$ and $date_2$ is 9. For this example, the combinations $\mathbf{m_1{:}p_2{:}w_2{:}s_4}_{clicks}$ and $\mathbf{m_1{:}p_2{:}w_2{:}s_6}_{clicks}$ did not occur during the specified period of time and thus, were not included in either **Q** or **Q'**. However, they might have appeared if a larger time range had been specified. In our implementation, we initialize **M:P:W:S(m,p,w,*)** as an empty set and update it each time an event occurs in the system. Since adding or checking whether a member exists within a set is a constant time operation, this does not degrade the performance of the system. In this way, **M:P:W:S(m,p,w,*)** only contains combinations for which data is available, which in turn allows us to efficiently perform the intersection operation.

B. Approach 2

A simpler and faster alternative to the previous approach would be storing the desired combination of keys to answer the report query in a unique sorted set as well. In that case we would not need to perform any intersection operation over ad spaces but only set unions over time. Thus, for the previous example, if we stored a-priori the sorted set "**M:P:W**$(time)$" with the corresponding values "**m:p:w**$_{clicks}$" and their scores inside, we could answer the

 ISBN: 978-1-61208-242-4

previous question by simply performing a union operation over the time period as follows:

$$\mathbf{Q'} = \bigcup_{d=date_1}^{date_2} \mathbf{M:P:W}(d) \quad (7)$$

As the union operation automatically performs an addition of the scores, after it finishes we just need to look at the score associated to the "**m:p:w**$_{clicks}$" combination we are interested in. This approach is considerably faster and simpler, but it is more expensive in terms of resources as more sorted sets need to be created.

C. Approach 3

A third approach to deal with this type of queries is simply naming the sorted sets, not by using generic dimension names as done by the first two approaches, but instead by naming them using the individual ids of the dimensions of interest and storing the facts in them. Hence, in the case of the above example, a sorted set could be named as **M:P:W**$(\mathbf{m}, \mathbf{p}, \mathbf{w}, time\_slot)$, and in it we can store "clicks (with score), views (with score)", where the score in each value represents the total number of clicks and views, respectively, accumulated during the period of time defined by time_slot for that unique combination of dimension ids (**m** and **p**). Furthermore, in this case it makes even more sense to keep only the keys of the deep-most branches and their combinations, as the less granular queries can be answered by performing simple union operations over their child sorted sets:

$$\bigcup_{d=date_1}^{date_2} \mathbf{M:P:W}(\mathbf{m}, \mathbf{p}, \mathbf{w}, d) \quad (8)$$

The main difference with the first approach is that in this case it is not necessary to look for the specific combination of keys inside the resulting set but this output set will readily provide the value of interest. The disadvantage is that it creates considerably more sorted sets than in the first two approaches.

D. Non-aggregate events

In all of the above examples, sorted sets and union operations work well since clicks and views correspond to events that are aggregate statistics. But, this is not true for all the cases, like unique users for instance. So, how can we answer a query like: "what is the number of unique users that clicked on campaign resource **r** from campaign **c** ran by advertiser **a** between $date_1$ and $date_2$?". We cannot simply add the number of daily unique users throughout the whole comprehended period, because a given unique user could access the system at two or more different time slots. For this case, we do not actually need neither sorted sets nor scores but normal sets defined, for instance, as in Approach 3, with the only difference that instead of the basic keys, we can

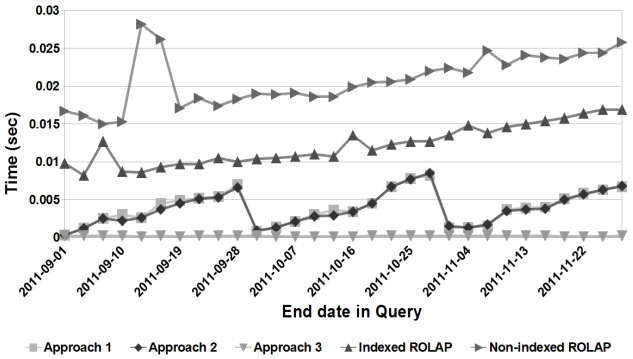


Figure 3. Performance comparison for Query #1.

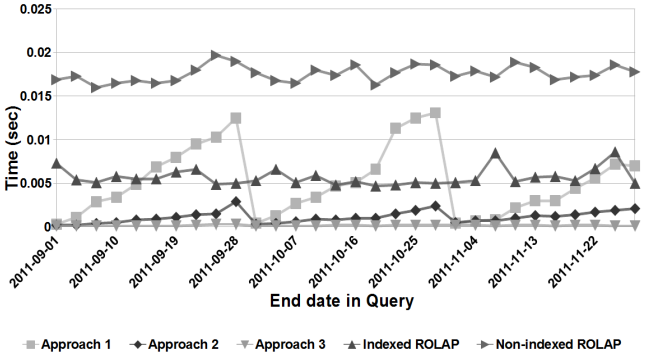


Figure 4. Performance comparison for Query #2.

directly store the non-aggregate statistic of interest (user_id in this case) inside sets like the one described below.

$$\mathbf{Users:A:C:R}(\mathbf{a}, \mathbf{c}, \mathbf{r}, time\_slot) = \{user_1, \ldots, user_n\} \quad (9)$$

The example set above stores the unique users detected per time_slot (1 hour in this case) for the combination campaign resource **r**, campaign **c** and advertiser **a**. Thus, whenever there is a new view for that combination of dimension ids, we simply store the user_id into the set. If the user_id is already an element in the set then no change takes place. If the user_id is not an element in the set, it is added to it. To answer the example query stated above we only need to get the cardinality of the resulting set **U** from the union operation over the whole time period as stated below:

$$|\mathbf{U}| = \left| \bigcup_{d=date_1}^{date_2} \mathbf{Users:A:C:R}(\mathbf{a}, \mathbf{c}, \mathbf{r}, d) \right| \quad (10)$$

V. Tests and Results

The proposed system was implemented on Redis 2.4.2 for a real ad server application running on Ruby on Rails 2.3.10 and MySQL [3]. The models previously used as example in this paper come from the same platform but have been pruned and much simplified for the sake of clarity.

 ISBN: 978-1-61208-242-4

A. Experimental settings

To assess the performance of our three proposed approaches, we present a set of seven sample queries for our ad server application, which cover many relevant cases involving multidimensional data so that other specific queries can be easily derived from them.

- **Query 1:** What is the total number of clicks and views between any two dates $date_1$ and $date_2$ for campaign resource **r** from campaign **c** owned by Advertiser **a** and Manager **m**?
- **Query 2:** What is the total number of clicks and views between any two dates $date_1$ and $date_2$ on the advertising space **s** owned by Publisher **p** and Manager **m** when shown on applications?
- **Query 3:** What is the number of clicks and views between any two dates $date_1$ and $date_2$ for campaign resource **r** from campaign **c** owned by Advertiser **a** when shown in web sites on the ad space **s** owned by Publisher **p** and Manager **m**?
- **Query 4:** What is the number of clicks and views between any two dates $date_1$ and $date_2$ for campaign resource **r** from campaign **c** owned by Advertiser **a** when shown on all ad spaces of all websites owned by Publisher **p** and Manager **m**?
- **Query 5:** What is the number of clicks and views between any two dates $date_1$ and $date_2$ for the whole campaign **c** owned by Advertiser **a** when shown in applications on the ad space **s** owned by Publisher **p** and manager **m**?
- **Query 6:** What is the number of unique users that clicked on any campaign resource from campaign **c** owned by Advertiser **a** when shown on applications in ad space **s** owned by Publisher **p** and Manager **m**?
- **Query 7:** What is the number of unique users that clicked on all campaign resources owned by Advertiser **a** when shown on all ad spaces owned by publisher **p** and manager **m**?

The data used for all queries correspond to real production data taken from September 1st, 2011 through November 30th, 2011, i.e. exactly three months.

In a first performance measurement attempt, an open-source ROLAP database based on a *Pentaho's Mondrian* server [4] running on top of the MySQL database holding the ad server application's data was set up and tested. Its performance, however, was rather slow when answering the seven example queries taking in average between one and three minutes. In view of the situation, an optimized ROLAP database was setup by providing the underlying MySQL database with pre-computed tables holding all the measures of interest (clicks and views) per hour and, furthermore, either provide or not the MySQL data with tailored indexes to answer each of the seven queries. This optimized MySQL database provided to Mondrian is denoted as *optimized ROLAP* and it is the benchmark against which our approach is compared.

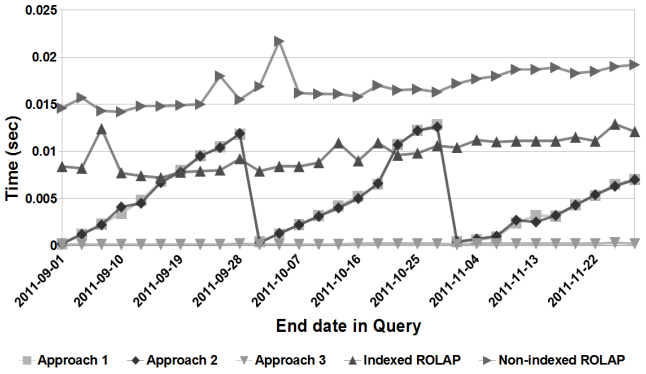


Figure 5. Performance comparison for Query #3.

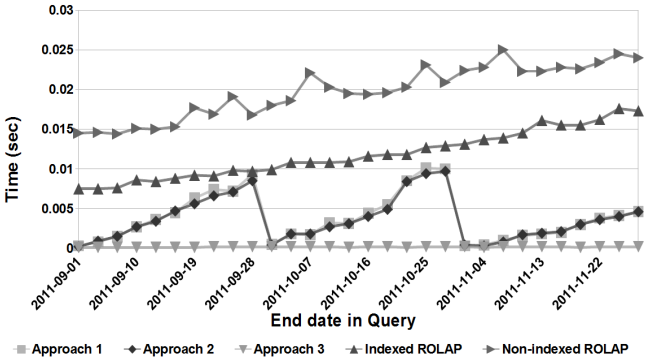


Figure 6. Performance comparison for Query #4.

One of the most relevant Redis functionalities is the policy for automatic deletion of keys in the database. Our system should only keep sorted sets that are frequently consulted or queried, and at the same time automatically remove those that are seldom seen or used. Fortunately, Redis provides a set of policy options to automatically remove keys so that the amount of memory allocated to the databases is never exceeded. In our implementation we have set the *least used resource* policy so that when the system reaches its maximum allocated memory, it can start removing the least used sorted sets from the database.

B. Experimental Results

Briefly, the first five queries (Figures 3 through 7) correspond to aggregate statistics (clicks and views) while the last two (Figures 8 and 9) are associated with non-aggregate statistics (unique users). For the former five the performance of the three approaches described in Sections IV-A, IV-B and IV-C is measured and compared with the *optimized ROLAP* (with indexed and non-indexed MySQL data) with pre-computed tables holding all the measures of interest per hour. For the last two queries, the performance of the non-aggregate statistics method described in Section IV-D is also

 ISBN: 978-1-61208-242-4

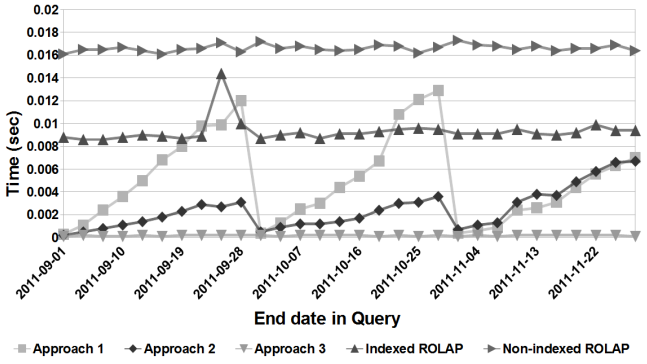


Figure 7. Performance comparison for Query #5.

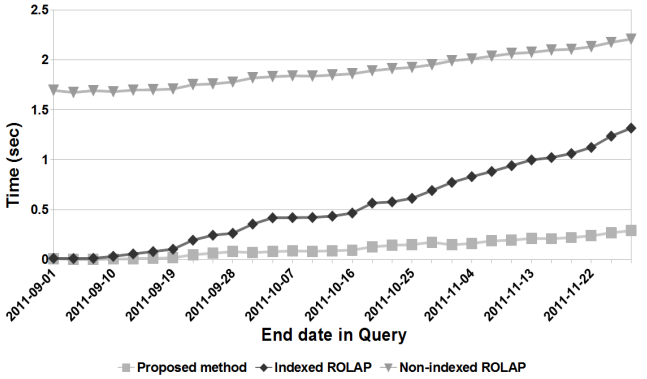


Figure 8. Performance comparison for Query #6.

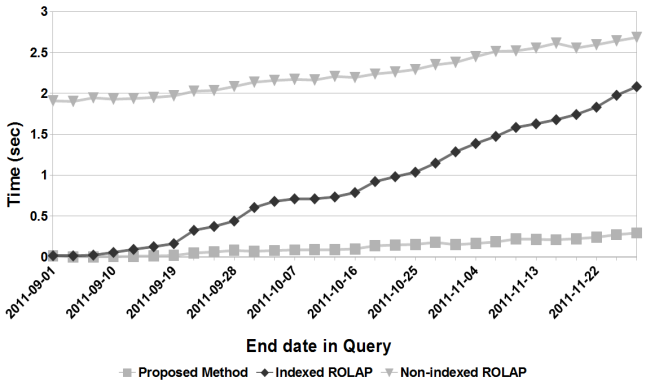


Figure 9. Performance comparison for Query #7.

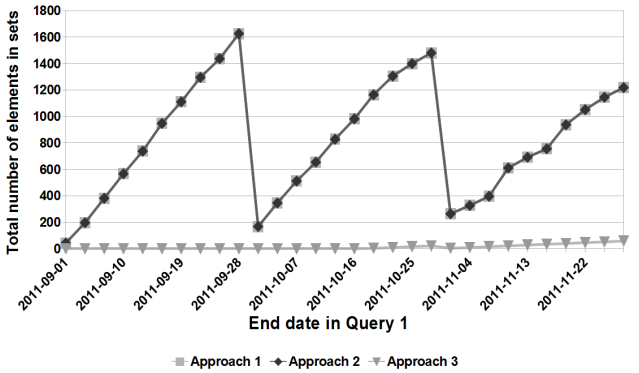


Figure 10. Number of elements in sets for Query 1.

benchmarked against the same *optimized ROLAP*.

The maximum resolution chosen for our application is one hour. Thus, the sorted sets of the three proposed approaches are named using the time dimension elements hour, day, month and year. That means we can always access pre-computed aggregate data in those time scales for any combination of dimension ids used in our sorted sets. From Figures 12 and 13, one can see that the ratio among the three proposed approaches in terms of both memory-usage and number of created sorted sets in Redis is roughly 1:3:6, for approaches 1, 2 and 3, respectively. Regarding the execution time for the seven sample queries, first of all it is important to remark that the abrupt falls in the execution time every 30 days are due to the use of our pre-computed aggregate monthly data in our system. It can be observed in the results that our proposed approaches 2 and 3 are faster than appropriately indexed MySQL in all cases except in the case of Query 3. This is because Query 3 corresponds to the most granular query in our system as it involves all single nodes in the ORM from the root to the bottom leafs in all branches. As such, it is required to perform a union operation on large sets of per-day aggregate data with hundreds of elements each. Despite this, we can see that, even in case of Query 3, our system outperforms *properly indexed optimized ROLAP* most of the times. In the context of this paper, *properly indexed* means with indexes tailored for the incoming queries.

It is interesting to observe that in the case of queries 6 and 7 for non-aggregate statistics, the proposed approach is more than 10 times faster than non-indexed *optimized ROLAP*, and 5 times faster than indexed *optimized ROLAP* for three months of data. For this approach, however, the system needs to create a large amount of sets, making the system inefficient from a memory usage point of view. Figures 10 and 11 show the total number of set elements for approaches 1, 2 and 3 when answering the first and second queries. As it can be seen in Figure 10, the number of elements in approaches 1 and 2 is exactly the same for Query 1, but substantially differ in Query 2, as shown in Figure 11. This is because Query 1 involves a combination of dimensions directly included in the basic set of keys, while for answering Query 2 new sets must be created in Approach 1. These new sets are created by union operations over the basic-key sets, as the combination of dimensions requested in the query involves only partial branches from the relational tree. This also explains the execution time of Approach 1 being considerably larger than Approach 2 in queries 2 and 5 (Figures 4 and 7, respectively).

 ISBN: 978-1-61208-242-4

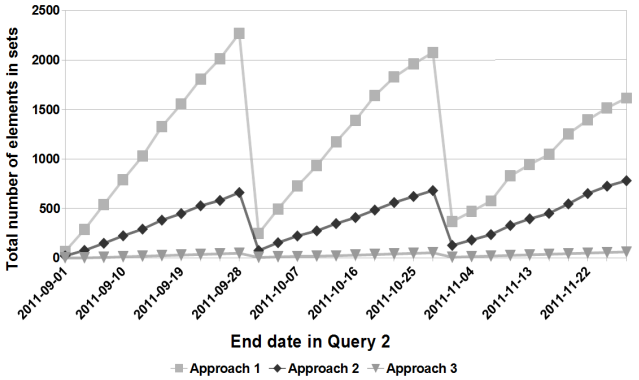


Figure 11. Number of elements in sets for Query 2.

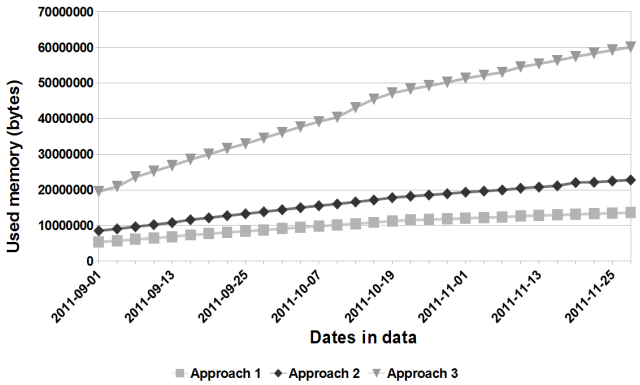


Figure 12. Memory usage.

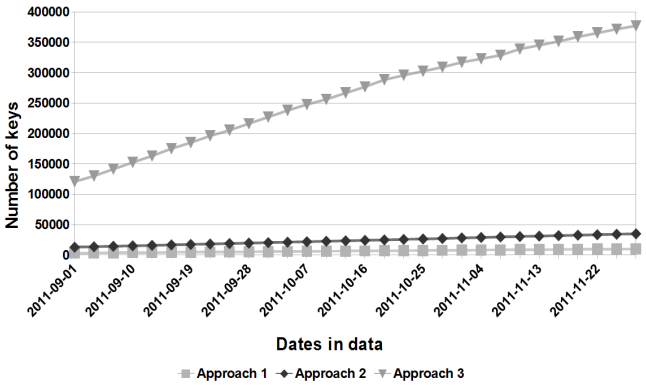


Figure 13. Number of keys.

C. Discussion of Results

As expected, the results show that Approach 3 is the fastest among all three (in all cases its execution time falls between 100 and 200 microseconds), but at the same time it is very inefficient in terms of memory: it consumes thrice as much as Approach 2 and six times as Approach 1. On the other hand, Approach 2 uses only twice the memory of Approach 1 but significantly outperforms it in terms of query execution time for all sample queries. Hence, Approach 2 is a much better performance compromise than Approach 1 and a very good alternative in comparison with Approach 3 when the memory usage is a big issue. That does not mean, however, it is impractical to use Approach 3 or Approach 1 in all kind of scenarios since as it can be observed in Fig. 12 the consumption of memory increases linearly with the amount of stored data for the three approaches. Thus, Approach 3 can be used in applications that feature a small number of RDB tables - i.e. a small number of sorted sets in Redis - with a lot of data inside while Approach 1 becomes attractive in applications where very granular queries are common since that means full-path queries in the proposed Redis data structure would be used more often.

Our results show that the speed of MySQL considerably improves after appropriately indexing the database on different dimensions to match those of the concerning query. However, it is well known that it is inefficient to keep many composite indexes especially on databases subject to frequent INSERTs, UPDATEs and DELETEs (as in our case) and furthermore, it is impractical to keep composite indexes for all possible multidimensional queries. Problems related to multi-column indexes in relational databases are well documented in the literature [11], but these topics are beyond the scope of this paper. A big advantage of the proposed system is that it does not have to pre-compute any sorted set but it creates them on the fly as the queries arrive following one of the described approaches. That means no further delay is introduced by our system in order to build the sorted sets as it works on the caching layer from the application's viewpoint and no data has to be transferred from the RDB to the proposed system in a cold start.

VI. Conclusions and Future Work

This paper has proposed and analyzed a new method to map a relational database into sorted sets on a key-value store database in order to create OLAP-like reporting systems. The four approaches described to map relational databases and their OLAP facts into sorted sets in key-value store databases for both aggregate and non-aggregate statistics are easily applicable to any conventional RDB structure. Our implementation in Redis shows that the proposed system is fast, surpassing properly indexed MySQL, to answer complex multidimensional queries for both aggregate and non-aggregate data. Furthermore, the flexibility of the database schema and the presence of automatic key deletion policies make the implementation very efficient. In our view the proposed scheme opens an innovative way to deal with complex data visualization and reporting systems based on multidimensional data using key-value store databases. Future work will aim at better algorithms to tune the system, studying alternatives for distributed systems based on other key-value data stores and reduce the amount of memory used by sets related to non-aggregate data.

References

[1] Druid. [Accessed Jul 13, 2012] http://www.metamarkets.com.

 ISBN: 978-1-61208-242-4

[2] Java programming language. [Accessed Jul 13, 2012] http://www.java.com.

[3] Mysql. [Accessed Jul 13, 2012] http://www.mysql.com.

[4] Pentaho mondrian project. [Accessed Jul 13, 2012] http://mondrian.pentaho.com.

[5] Python programming language. [Accessed Jul 13, 2012] http://www.python.org/.

[6] Qlikview. [Accessed Jul 13, 2012] http://www.qlikview.com.

[7] Redis. [Accessed Jul 13, 2012] http://www.redis.io.

[8] Ruby programming language. [Accessed Jul 13, 2012] http://www.ruby-lang.org.

[9] What is object/relational mapping? [Accessed Jul 13, 2012] http://www.hibernate.org/about/orm.

[10] D. J. Abadi. Data management in the cloud: Limitations and opportunities. *IEEE Data Eng. Bull.*, 32(1):3 – 12, 2009.

[11] D. J. Abadi, S. R. Madden, and N. Hachem. Column-stores vs. row-stores: how different are they really? In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 967 – 980, New York, NY, USA, 2008. ACM.

[12] A. Abelló, J. Ferrarons, and O. Romero. Building cubes with mapreduce. In *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, DOLAP '11, pages 17 – 24, New York, NY, USA, 2011. ACM.

[13] G. Alonso, D. Kossmann, and T. Roscoe. Swissbox: An architecture for data processing appliances. In *CIDR*, pages 32 – 37. www.crdrdb.org, 2011.

[14] L. Bonnet, A. Laurent, M. Sala, B. Laurent, and N. Sicard. Reduce, you say: What nosql can do for data aggregation and bi in large repositories. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pages 483 – 488, September 2011.

[15] P. Brezany, Y. Zhang, I. Janciak, P. Chen, and S. Ye. An elastic olap cloud platform. In *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*, pages 356 – 363, December 2011.

[16] Y. Cao, C. Chen, F. Guo, D. Jiang, Y. Lin, B. C. Ooi, H. T. Vo, S. Wu, and Q. Xu. Es2: A cloud data storage system for supporting both oltp and olap. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, pages 291 – 302, April 2011.

[17] R. Cattell. Scalable sql and nosql data stores. *SIGMOD Rec.*, 39:12 – 27, May 2011.

[18] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.*, 26:4:1 – 4:26, June 2008.

[19] C. Chen, G. Chen, D. Jiang, B. C. Ooi, H. T. Vo, S. Wu, and Q. Xu. Providing scalable database services on the cloud, 2010.

[20] E. F. Codd, S. B. Codd, and C. T. Salley. Providing olap to user-analysts: An it mandate. *Ann ArborMichigan*, page 24, 1993.

[21] A. Cuzzocrea, I.-Y. Song, and K. C. Davis. Analytics over large-scale multidimensional data: the big data revolution! In *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, DOLAP '11, pages 101 – 104, New York, NY, USA, 2011. ACM.

[22] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.*, 41:205 – 220, October 2007.

[23] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37:29 – 43, October 2003.

[24] R. Kallman, H. Kimura, J. Natkins, A. Pavlo, A. Rasin, S. Zdonik, E. P. C. Jones, S. Madden, M. Stonebraker, Y. Zhang, J. Hugg, and D. J. Abadi. H-store: a high-performance, distributed main memory transaction processing system. *Proc. VLDB Endow.*, 1:1496 – 1499, August 2008.

[25] D. Kossmann, T. Kraska, S. Loesing, S. Merkli, R. Mittal, and F. Pfaffhauser. Cloudy: a modular cloud storage system. *Proc. VLDB Endow.*, 3:1533 – 1536, September 2010.

[26] A. Lakshman and P. Malik. Cassandra: a decentralized structured storage system. *SIGOPS Oper. Syst. Rev.*, 44:35 – 40, April 2010.

[27] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. Madden, and M. Stonebraker. A comparison of approaches to large-scale data analysis. In *Proceedings of the 35th SIGMOD international conference on Management of data*, SIGMOD '09, pages 165 – 178, New York, NY, USA, 2009. ACM.

[28] L. Qin, B. Wu, Q. Ke, and Y. Dong. Saku: A distributed system for data analysis in large-scale dataset based on cloud computing. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 2, pages 1257 – 1261, July 2011.

[29] Y. Wang, A. Song, and J. Luo. A mapreduce merge-based data cube construction method. In *Grid and Cooperative Computing (GCC), 2010 9th International Conference on*, pages 1 – 6, November 2010.

 ISBN: 978-1-61208-242-4

Ontology-Guided Data Acquisition and Analysis

Using Ontologies for Advanced Statistical Analysis

Dominic Girardi, Michael Giretzlehner
Research Unit Medical Informatics
RISC Software GmbH
Hagenberg, Austria
firstname.lastname@risc.uni-linz.ac.at

Klaus Arthofer
Department for Process Management in Healthcare
University of Applied Sciences Upper Austria
Steyr, Austria
klaus.arthofer@fh-steyr.at

***Abstract*—We present an ontology-based, domain independent data acquisition and preparation system, which is able to process arbitrarily structured data. The system is fully generic and customizes itself automatically at runtime based on a user-defined ontology. So it can be instantiated for any domain of application by defining an ontology for this domain. Furthermore, semantic rules can be integrated into the ontology to check the data's semantic plausibility. We plan to integrate statistical and data mining algorithms that take advantage of the structural ontology information to allow the user to perform (semi) automatic explorative data analysis. In this paper, the system is described in detail and a motivation for ontology-guided data analysis is given.**

Keywords - ontology-based data acquisition; semantic data quality; ontology-supported data analysis.

I. INTRODUCTION

Data analysis is one of the last steps in the process of a data-based research project. Surely, it is the most important one; the one that extracts the information out of the data which has been collected before. Still, the quality of the outcome is limited by the quality of the preceding steps, which are: data acquisition, data validation and cleaning, and data preparation. While syntactical data validation is well established, checking for data semantics is widely neglected; despite the fact that the problem of bad semantic data quality is serious and well known. Semantic data quality matters the data validity concerning the meaning and no syntactical aspects [19]. Using data of insufficient semantic data quality has fatal consequences regarding the usability of reports – keyword garbage in, garbage out. Consequently, data analysis cannot be considered as an isolated working step, but has to be integrated into a process containing all other steps, mentioned above.

We present an ontology-based, generic, web-based data acquisition and preparation system, which is able to store and process data of arbitrary structure and is therefore applicable to any domain of application. By modelling the domain of application as an ontology, the domain expert prepares the system for data acquisition. The rest of the system, including web-based user interfaces for data acquisition and import interfaces for importing electronically stored data, are created automatically at runtime, based on the ontology information. Furthermore, the system allows the definition of semantic plausibility rules to ensure not just the syntactic correctness of the data but also the semantic one. The formalized domain knowledge, which exists in the form of an ontology, is going to be used to guide and configure statistical analysis algorithms on the data. This allows the domain experts – who are most likely no IT experts – to set up and run their own data acquisition system without the need for an IT or database expert.

Section 2 contains an overview over related research projects. In Section 3, we provide a motivation and the theoretical background for our generic data acquisition and semantic checking infrastructure and provide key numbers of already running data acquisition projects. In Section 4, we describe that kind of statistical analysis features we want to integrate into the system and how the ontology helps to improve the results of the analysis. Our conclusions can be found in Section 5.

II. RELATED RESEARCH

Ontologies are widely used for flexible data integration, where data from multiple heterogeneous sources is mapped into a central ontology. In their one page position paper Zavaliy and Nikolski [12] describe the basic concept of an ontology-based data acquisition system for electronic medical record data. They use a very simple ontology, which contains four concepts (*Person*, *Hospital*, *Diagnosis* and *Medication*). The *Web Ontology Language* (OWL) [20] is used for modelling their domain. They point out, that the main reason for using an ontology-based approach is the need for adaptive data structures. This work is closely related to our work. However, their paper contains no information on how the data can be entered into the system, neither is there information about the system architecture or semantic data checking. There is also no information given on how adaptable their ontology is and if those four main concepts can be replaced or not. Despite the fact that they follow a very similar basic idea, our system is more extensive and matured. Guo and Fang [13] describe an ontology-based data integration system. They use ontologies to cope with the semantic and structural heterogeneity of data from different source applications. This is closely related to one aspect of our system, namely the automatically created data import interfaces. They can be used to import data from

 ISBN: 978-1-61208-242-4

heterogeneous data sources. While data import is only one aspect of our system, Guo and Fang [13] focus on this matter and provide sophisticated concept mapping algorithms to improve the integration process. In their proposal, Dung and Kameyama [14] describe an ontology-based health care information extraction system. They modelled the medical domain with an ontology and use the semantic information of this ontology to extract information from texts. A so called "*New Semantic Elements Learning Algorithm*" extends the ontology semi-automatically. Although their approach differs in some ways from ours (information extraction vs. data acquisition and analysis), both share a common idea: Information from heterogeneous systems is stored into a central ontology. Nevertheless, their ontology is very closely related to their field of application, while our system is fully domain independent.

Lin et al. [15] provide an ontology for data mining processes. The knowledge, which is stored in the ontology, helps the user to decide which data mining algorithm should be used for the task on hand. Although we use the ontology primarily for data definition, the integrated data mining algorithms need to know how to treat the current data. This information can be derived from our ontology. So, while Lin et al. [15] use an explicit ontology for guiding the data mining process, our ontology will guide the process more implicitly be providing meta-information about data types, etc. Zheng et al. [16] identified the a gap in data mining processes that arises between technicians, who perform data mining, and the domain experts, who's knowledge is needed to interpret and guide it. They use two ontologies to cope with this problem: a domain-ontology, where domain experts enter their knowledge, and a task-ontology for choosing the best data mining algorithms for the current problem. While the latter one is closely related to [15], the first one tries to enforce the connection between technicians and domain experts. Viewed in this light, they try to solve the same problem as we do in the exactly opposite way. While they want to support the technician with domain knowledge, we want to enable the domain expert to run data mining algorithms. Nimmagadda and Dreher [17] give a good example how ontologies are used in praxis. They modeled the complex domain of oil production using an ontology to support the data integration and mining process.

Despite the fact that the concept of building an application upon an ontology is widely used, we could not find any system that implements this concept as consequently as we do. In most cases the systems are closely connected to their field of application and the domain in the background is not arbitrarily exchangeable. Furthermore, we could not find any ontology-based systems, which cover all aspects from data acquisition (manual via automatically created web interface and electronically via generated interfaces) to data storage, semantic data checking and data analysis.

III. ONTOLOGY-BASED DATA ACQUISITION

A. Technical Background

Whenever data of a non-trivial structure is collected for statistical analysis, a professional data acquisition and storage system is needed. Due to the semantic dependency of the systems' data structures from the domain of application they are usually inflexible and hardly reusable for other domains.

To overcome this drawback and resolve the dependency, we developed a data acquisition and storage system, which is not based upon a domain-specific data model, but on a generic meta-data model. The meta-data model is able to store the actual data model the form of an ontology.

According to the definition of Chandrasekaran et al. [1] "*Ontologies are content theories about the sorts of objects, properties of objects, and relations between objects that are possible in a specified domain of knowledge.*". Gruber [2] provides a more general definition: "*An ontology is a specification of a conceptualization.*". Technical details on the meta-model can be found in [18].

In this way, the system's meta-data model remains constant and independent from the domain of application. The process of defining the domain-specific ontology in the generic meta-data model is called instantiation of the generic meta-model by a domain specific ontology. This instantiation is performed by a domain expert with support of the Ontology Editor (see Section III.B). Furthermore, the collected data itself is stored into this meta-model.

The user interfaces for data input, including input forms, overview tables, search and filter functionality is automatically created at runtime, based on the ontology definitions. Numerous configurable properties ensure the possibility to customize the automatically generated graphical user interface (GUI).

B. System Architecture

The main purpose of the project is to create a system, which allows non-IT experts to set up and maintain their own professional data acquisition system for, e.g., research, clinical studies or benchmarking purposes. Furthermore, the system is able to store domain specific knowledge in terms of rules, in order to check the semantic quality of the stored data. The main parts of the system are:

- The Ontology Editor: This software allows the user to instantiate the generic meta-model with the ontology of his field of application.
- The Web Surface: The automatically created web surface allows the data collectors to enter their data into the system.
- The Semantic Check Engine: This part of the system checks each data record against the rules, defined by the user, to ensure its syntactic and semantic integrity.

In the following sections, important parts are described in detail.

C. Ontology Editor

The Ontology Editor allows the definition and maintenance of the ontology. The domain-expert defines

 ISBN: 978-1-61208-242-4

which data elements (classes) exist in his project, which attributes they contain, and how there are related to each other. Furthermore, the data type of each attribute has to be defined. The user can choose between text, integer, float, numerous date formats, and enumeration types. The latter ones are displayed as lookup tables. In order to keep large enumerations applicable they can be organized in hierarchical structures, resulting in taxonomies of enumerations. The ontology can be changed and adapted at any time.

Moreover, the Ontology Editor allows the display of the stored data sets and offers numerous filter, search and batch processing functions for administrating large data sets. Since the structure of the data depends on the actual ontology, all GUI structures (tables, headers, filter dialogs, etc.) are created at runtime, based on the ontology information.

D. Semantic Data Check

A benefit of ontology-based data-acquisition systems is the possibility to allow the domain-expert to manage enumeration types, without being dependent from an IT company to overtake this task. Due to the extensive use of adaptable enumerations the demand for free text input field is reduced to a minimum. Free text fields are the worst case for automatic processing for both: semantic checking and statistical analysis. Furthermore, the extensive use of enumerations forces the domain-expert to maintain these enumeration in order to keep them up to date and if the number increases – to organize them in meaningful hierarchies. This represents a central aspect of master data management.

For defining the rules for semantic data checks, the user has to establish so-called dependency relations (short: dependencies) among two processible attributes. Processible in this case means: not a free text attribute. One of the attributes is the master attribute; the other one the slave attribute. As the names suggest, the value of the master attributes defines the plausibility of the value of the slave attribute. In other words, the plausibility of the slave attributes depends on the value of the master attribute. E.g., the master attribute is the attribute *Gender* of the class *Patient*, and the slave attribute is the attribute *Diagnose* of the class *Disease*, then the diagnose *Pregnancy* is not plausible if the gender is *Male*. If one slave is controlled by more than one master, then these dependencies need to be connected by a logical operator AND or OR. This results in a logical tree of conditions, which is processed to determine the semantic plausibility of the given data set. Fig. 1 shows the configuration interface for this given example. The slave element (on the left hand side) is *Diagnose*, whereas the dependency is only set for the diagnose *Pregnancy*. The list on the right shows all possible master enumeration values for the master *Gender* (male and female). The state of the checkboxes indicate that a diagnose *Pregnancy* is plausible if the gender is *female* and not *male*. Asides from the actual configuration it shows the logical expression tree in which this dependency is embedded.

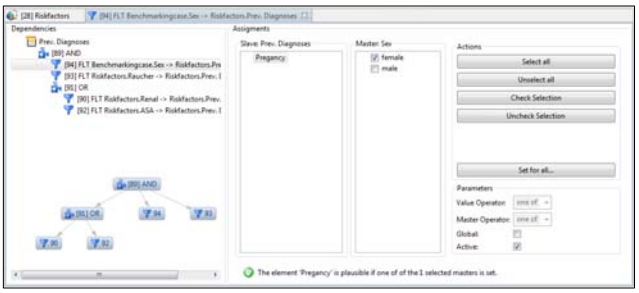


Figure 1. Configuration dialog for a new dependency

One single attribute can be master of one dependency and at the same time slave of another dependency. So transitive dependency graphs can be created, which are processed from the leaves to the root of the tree using the constraint propagation scheme [3].

Before the data is used for statistical analysis, these checks are performed, and a detailed report is created. Check results are listed and the conflicts are described in detail. Combined with the check for syntactical correctness and the checks for static constraints (minimum and maximum values) the result of the semantic check provides a quality report of the current data set.

E. Experiences and Key Numbers

Currently, the described infrastructure is used to perform comparative benchmarking of surgical treatments of hospitals in Upper Austria. Each quarterly period the data is imported electronically from heterogeneous hospital information systems and supplemented with handwritten patient information by specially trained study nurses using the automatically generated web interface.

For each benchmarking cycle (3 months) about 1,500 medical cases are entered into the system, containing medical information as well as administrative data of a patient's treatment in hospital. An average case contains about 50 data elements (diagnoses, treatments, etc.), which results in about 75.000 data elements for a benchmarking cycle. Fig. 2 shows the automatically created web interface for this project. It displays one particular medical case. The tree on the left hand side visualises the whole case including all data elements, whereas its structure is derived from the ontology. The input form in the middle is also dynamically created based on ontology information and allows the editing of the currently opened case. The red fields on the right present the result of the semantic data check. While the colour indicates the result, a detailed list of errors is shown, when the user clicks on the field.

Before they are statistically analysed, semantic checks are performed. First runs of Semantic Check Engine emphasized the importance of semantic data checks and showed high error rates. Compares between manual and automated semantic checks showed a time saving of up to 15 minutes per medical case (about 15 minutes for manual checking by a domain expert vs. about 20 seconds for automated checking), resulting in an overall saving of more than three work weeks.

First clinic results are about to be published this year.

 ISBN: 978-1-61208-242-4

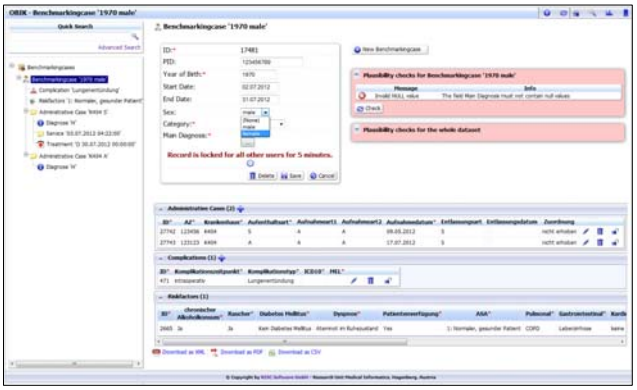


Figure 2. The automatically created Web Interface

IV. ONTOLOGY-GUIDED DATA ANALYSIS

The possibility to embed statistical analysis algorithms directly into the system offers numerous advantages for data analysts. We plan to extend our data acquisition system by a set of statistical analysis and data mining algorithms that enables analysts to get a quick overview of the data. The ontology information allows the (semi) automatic data mining, and helps to reduce the human bias on statistical analysis.

A. Ad hoc Analysis

In many cases, data acquisition and storage is done by different systems than data analysis. The data has to be extracted from the first and imported into the latter one. Although state-of-the-art data analysis systems like SAS, SPSS, etc., offer a big variety of data import interfaces, this process takes time. In the event that the most important statistical key features like descriptive statistic parameters, histograms, statistical hypothesis testing, etc., are directly integrated into the system, the analyst can answer simple questions right from within the system. The integrated filter and search functionalities help to extract the datasets of interest and automatically compute the most important key numbers. The data type of each attribute, which is provided by the ontology, helps to interpret the data and calculate the correct key features.

Moreover, correlations between all attributes and structural features (e.g., number of sub-elements of a certain class) can be computed automatically and all relevant results are presented to the user. In that way, unexpected correlations can be discovered, which reduces the human bias to statistical analysis.

B. Semantic Data Quality

As we motivated in Section III.D, semantic data checks help to increase the quality of statistical analysis. But there's a way to return this benefit in a way, that data analysis helps to improve data quality. Assume a strong linear correlation between two attributes a_1 and a_2. Consequently, records that show a configuration of a_1 and a_2 that sheers out of this correlation are suspicious and can be proposed for addition investigation.

Prototype tests on a data set of biometric measurements of over 1500 children showed very strong linear correlations between the different measurements of the human body. Several statistical outliners from these correlations were detected and revised. Although their actual data was within the defined ranges, these errors could be identified because their combination was implausible. Other tests showed that the weight of birth of children distributed normally – which we expected. More than ten input errors could be identified because of the significantly high distance of the data set from the mean of the distribution. With these two tests, we could show that even very simple statistical analysis can help to increase the quality of the stored data.

C. Using Taxonomies for Higher Level Analysis

The features described so far help to reduce the effort of data analysis by automation. For the following feature, the strong connection between the ontology and the analysis is essential.

Dealing with large enumerations quickly leads to sparse data sets. For examples, when dealing with medical data the diagnosis of a patient is often defined using the ICD-10 (International Classification of Diseases) catalogue, which consists of more than 12,000 entries. When, e.g., a Chi^2 [21] test is used to identify a correlation between the diagnoses and a treatment (coded with an enumeration type of a similar size), even with several thousands of data sets, the table would still contain lots of empty cells, and the result would be not be interpretable.

In this case, the hierarchically organized enumeration types help to reduce the number of rows and columns of the Chi^2 test setup. Since the hierarchical relations in the enumeration type is a IS-A relation the numerous enumeration values can automatically be summarized in meaningful groups; provided that the hierarchy was designed properly. So, instead of thousands of low level enumeration values less high level enumeration concepts are analysed; similar to the approach of Xiangdan et al. [4], where the concepts of ontologies are used to discover fewer high-level rules, instead of lots of low-level rules.

D. Data Visualization including Structural Features

One of main objectives of explorative data analysis is to get an overview of the data. A very popular instrument for this purpose is the self-organizing map (SOM), which was introduced by T. Kohonen [5]. A SOM maps an *n*-dimensional input space into an *m*-dimensional (usually m=2) output space, whereas it converts the nonlinear statistical relationships between high-dimensional data into simple geometric relationships [6]. Traditional SOMs work on numeric vectors for both input format and internal data representation. So, for using the SOM, the relational data has to be transferred into a numeric input vector. The ontology supports this transformation process in a way that allows persons, who are not necessarily IT experts, to perform this transformation just by choosing the attributes that shall be considered. Fig. 3 shows the result of the visualization of a medical dataset with a SOM.

However, the highly structured data has to be transferred into a one dimensional numeric vector, where most of the structural information is lost. In a further step, SOM algorithms for structured data (SOM-SD) will be evaluated

 ISBN: 978-1-61208-242-4

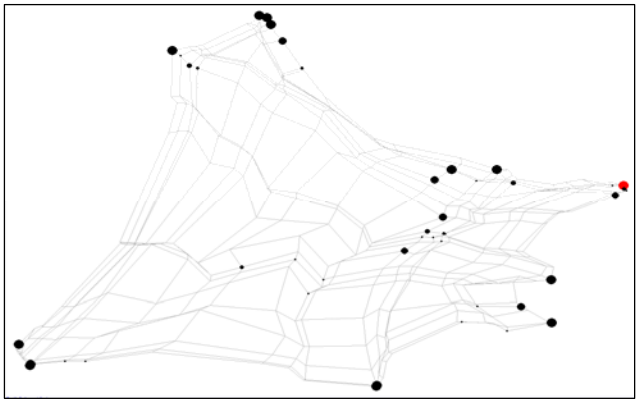


Figure 3. Data Visualization with a SOM

[7], [8], [9], [10]. For performing these algorithms access to the structural information is needed. So, this can only be automated and thus be made applicable for non-IT experts, by the strong linkage with the ontology.

V. CONCLUSION

In our research project, we could show that meta-model based systems help to reduce the setup effort of data acquisition and storage systems to a minimum. Additionally, the generic meta-model based approach allows the development of software features for classes of problem, not for just one single domain of applications. This guarantees a high degree of reusability of the system.

Furthermore, first semantic checks on real medical datasets showed high error rates concerning the sematic data quality in the tested hospitals, confirming what is conducted by Hüfner [11]. Without semantic data checks these error rates would influence the results of the analysis unnoticed, resulting in useless or suboptimal conclusions and consequences.

Statistical evaluation of the collected data is still performed in external systems, such as SAS, and SPSS, after exporting the data from the ontology based storage system. The planned statistical analysis features will never be able to replace these systems; neither is this our objective; but they provide an overview over the whole dataset before the analyst exports the data. Moreover, the strong linkage between statistical analysis and the ontology allows more sophisticated analysis with inclusion of structural and conceptional (enumeration type hierarchies) features. It also allows persons with limited IT skills to used highly complex algorithms like the SOM to explore their data.

REFERENCES

[1] B. Chandrasekaran, J. Josephson, and V. Benjamins. “What are ontologies, and why do we need them?” Intelligent Systems and their Applications, IEEE 1999, 14(1) pp. 20–26.

[2] T. R. Gruber. “A translation approach to portable ontology specifications”. Knowl. Acquis. 1993,5(2) pp. 199–220 .

[3] F. Rossi, P. van Beek, and T. Walsh. Handbook of Constraint programming. 1 edn. Elsevier, Amsterdam and Boston (2006)

[4] H. Xiangdan, G. Junhua, S. Xueqin, and Y. Weili. “Application of data mining in fault diagnosis based on ontology”, Proceedings of the Third International Conference on Information Technology and Applications (ICITA’05), 2005 IEEEes, Journal of Information Science, 2010 36:306

[5] T. Kohonen. “The self-organizing map”. Neurocomputing 21 1998, pp. 1-6

[6] T. Kohonen. Self-Organizing Maps – 3rd edn. Springer, 2001.

[7] M. Hagenbuchner, A. Sperduti, and A.C. Tsoi. “A self-organizing map for adaptive processing of structured data.” IEEE Transactions on Neural Networks 14(3) 2003, pp. 491-505

[8] M. Hagenbuchner and A.C. Tsoi. “A supervised training algorithm for self-organizingmaps for structures”. Pattern Recognition Letters 26(12) 2005 pp. 1874-1884

[9] M. Hagenbuchner, A. Sperduti, A.C. Tsoi, F. Trentini, F. Scarselli, and M. Gori. “Clustering xml documents using self-organizing maps for structures.” in: Workshop of the initiative for the evaluation of xml retrieval 2005, pp. 481 - 496

[10] M. Martin-Merino and A. Munoz. “Extending the SOM algorithm to non-euclidean distances via the kernel trick”. In Pal, N., Kasabov, N., Mudi, R., Pal, S., Parui, S., eds.: Neural Information Processing. Volume 3316 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2004), pp. 150-157

[11] J. Hüfner. „Datenqualität im Krankenhaus. Kostenvorteile durch ausgereifte Konzepte“, http://www.tiq-solutions.de/download/attachments/425996/Datenqualitaet-im-Krankenhaus_Jan-Huefner_08-2007.pdf (24.01.2012)

[12] T. Zavaliy and I. Nikolski. “Ontology-based information system for collecting electronic medical records data”. In: Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), 2010 International Conference on, p. 125.

[13] G. Fangyua and Y. Fang. “An Ontology-Based Data Integration System with Dynamic Concept Mapping and Plug-In Management”. In: Information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on, vol. 3, pp. 324–328.

[14] D. Tran Quoc and W. Kameyama.” A Proposal of Ontology-based Health Care Information Extraction System: VnHIES.” In: Research, Innovation and Vision for the Future, 2007 IEEE International Conference on, pp. 1–7.

[15] L. Mao-Song, Z. Hui, and Y. Zhang-Guo. “An Ontology for Supporting Data Mining Process”. In: Computational Engineering in Systems Applications, IMACS Multiconference on, vol. 2, pp. 2074–2077.

[16] Z. Ling, L. Feng, and G. Hui. “The research of ontology-assisted data mining technology”. In: Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on, pp. 285–288.

[17] S. L. Nimmagadda and H. Dreher. „Petroleum Ontology: An effective data integration and mining methodology aiding exploration of commercial petroleum plays”. In: Industrial Informatics, 2008. INDIN 2008. 6th IEEE International Conference on, pp. 1289–1295.

[18] D. Girardi, J. Dirnberger and M. Giretzlehner. “Meta-model based knowledge discovery”. In: Data and Knowledge Engineering (ICDKE), 2011 International Conference on, pp. 8–12.

[19] A. Kurz. Data Warehousing. Enabling Technology. mitp-Verlag (1999)

[20] D. L. McGuinness and F van Harmelen. Owl web ontology language overview: W3c recommendation (10 February 2004).

[21] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. Statistik – Der Weg zur Datenanalyse. 7th edn. Springer Heidelberg (2011), pp. 467–469

 ISBN: 978-1-61208-242-4

Analysis of Streaming Service Quality Using Data Analytics of Network Parameters

Jie Zhang
Dept. Computer Engineering
Kangwon National University
ChunCheon, Korea
zarg_1982@hotmail.com

Hwa-Jong Kim
Dept. Computer Engineering
Kangwon National University
ChunCheon, Korea
hjkim3@gmail.com

Doo-Heon Ahn
Dept. Computer Engineering
Kangwon National University
ChunCheon, Korea
arcarpe@daum.net

***Abstract*—Quality of multimedia streaming service depends on network parameters such as bandwidth, delay, jitter and packet loss rate. In order to effectively improve the Quality of Service (QoS), it is needed to know which parameter is dominant in deterioration of the service quality at the moment. In the paper, we investigate the interdependency among network parameters in terms of finding out which parameter is dominant in the quality deterioration. We also studied the sensitivity of parameters on the change of quality in an emulated communication network. For these purposes, we performed experimental tests on streaming video with 16 testers, and we used transformed 5-level values for each network parameter to imitate a Likert style evaluation for each variable. It is found that bandwidth and delay affect more on service quality than jitter or loss rate.**

Keywords-Data analytics; Network parameter; Quality of Service; Streaming service

I. INTRODUCTION

During the last decades, quality management of real-time communications has been widely studied in order to provide improved multimedia services [1] [2]. With rapid development in audio-visual technology and products, various types of TV-based multimedia services have been introduced including high definition TV (HDTV). Recently, along with the wide spread of high speed cellular networks and wireless LANs, mobile multimedia services also became common for the tablet PC or Smartphone users. The increased wireless multimedia service is known as the key source of network traffic and service dissatisfaction [3] [4] [5].

The purpose QoS control in network service is to satisfy users. To satisfy the users, we should consider many factors simultaneously such as contents searching time, download speed, screen size, and contents itself, beside the network parameters such as bandwidth, delay, error rate or jitter. However we cannot satisfy all the resources simultaneously, so it is needed to prioritize factors to support. For example, we may choose large screen for some movies, high bandwidth for high quality image, and low jitter for conversations. For a given quality level, we need to choose optimal combination of network resources.

In the paper, we analyze simulation test data in order to find which network parameter dominantly affected service quality at the moment. In other words, we want to understand the relationship between quality factors in streaming service. However, it is difficult to take into account various network parameters and human factors together in the analysis because it is difficult to extract correct relationship between user satisfaction and the factor such as screen size, search time, download speed exactly. Therefore, in the paper, we considered only four typical network parameters (bandwidth, delay, jitter, and loss rate). As a further study, we can extend the number of factors in the dependency analysis following the rationale proposed in the paper.

The paper is organized as follows. Chapter 2 introduces related works for QoS studies. Chapter 3 describes an experiment for QoS related user experiment and result analysis follows on chapter 4. Chapter 5 is for conclusion and further works.

II. RELATED WORKS

QoS measurement for multimedia services has been widely studied to find an optimized network environment [6]. Kostas E. Psannis, Yutaka Ishibashi and Marios G. Hadjinicolaou presented an approach for multimedia streaming services, which used priority including dedicated bandwidth, controlled jitter for video interactive services that require additional resources to provide differently encoded video [7]. The research showed how the encoded bit rate and its bandwidth give influence to the video quality. Liuming Lu, Xiaoyuan Lu, Jin Li monitored stream video quality by observing packet losses, and showed how the packet loss ratio affects the quality of video streaming services [8].

Most typical network parameters used in the QoS analysis are bandwidth, delay, loss rate and jitter [6]. Bandwidth is the most significant parameter in multimedia streaming service, delay can cause unsynchronized video/audio frames, packet loss would be the reason of video error, and jitter may cause frame bursting. However, it is rarely studied to consider the parameters together in order to find their dependencies. In the paper, we focused on finding the relative importance of the four parameters in streaming service.

III. SIMULATION MODEL

In the paper, we measured the quality levels for video streaming service via simulation. The simulation model is composed of three parts: the streaming server, network simulator and client PC. Figure 1 shows the simulation model used in the paper.

A. The streaming server

The streaming server provides video contents. A standard HD (1280x720 resolutions with 30 frames per second) video is sent to the clients through a simulated network.

 ISBN: 978-1-61208-242-4

B. The Network simulator

The Network simulator is used to emulate a real network. We used simulation package Shunra Cloud [9]. The package can change the bandwidth, delay, loss rate and jitter via software settings. We can choose any value of network parameters. For example, we can modify the channel capacity to have any bit rate, e.g., 10Mbps or 7Mbps. If the server is transmitting a 9Mbps video stream, the 10Mbps channel would be enough. However, if the bandwidth is set to be 7Mbps, then it will suffer a shortage of bandwidth. In simulation, we used various channel speed (bandwidth) ranging from, for example, 5Mbps to 10Mbps, and monitored the video quality with different bandwidth settings.

We also can choose any value for delay, loss rate and jitter for network emulation. With this scheme, we can have infinite number of sets for combinations (b, d, j, l) where b, d, j and l represent specific value of bandwidth, delay, jitter and loss rate respectively. In order to investigate the effect of each parameter to the video quality with a finite number of simulations, we need to minimize the number of parameter set. For this purpose, we chose discrete values for each b, d, j, and l.

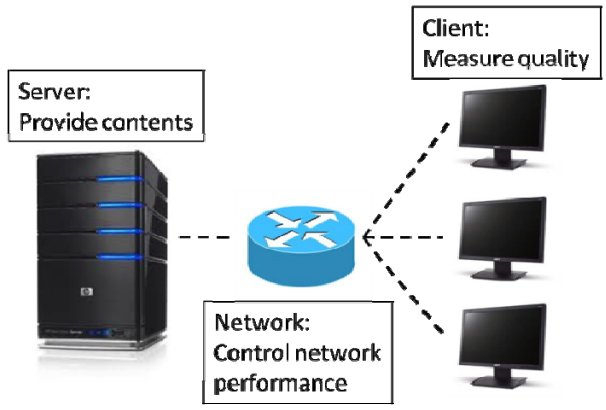


Figure 1. Simulation model.

In order to choose a reasonable finite number of parameter sets, we divided the range of each parameter into 5 quality levels (mimicking the Likert levels). For example, for the bandwidth parameter, first we set other parameter to have best conditions, that is, no delay, no error, and no jitter. Then we decrease the channel capacity (bandwidth) from 10Mbps down smoothly to find the Mean Opinion Score (MOS) evaluation to be 4 from 16 testers. Table 1 shows the 5 MOS threshold values for bandwidth, delay, loss rate and jitter. In the first column, when bandwidth is over 8.06Mbps, many people evaluated MOS 5, and with bandwidth of 7.85~8.06Mbps, many people evaluated the MOS to be 4, and so on. Below 7.47Mbps, many people evaluate MOS to be 1. In the evaluation we used the MOS such as, 5: Best, 4: Good, 3: Moderate, 2: Bad, 1: Worst.

TABLE I. LEVELS OF NETWORK PARAMETERS FOR EXPERIMENT

Likert	Bandwidth(Mbps)	Delay(ms)	Loss(%)	Jitter(ms)
5	8.06~	~257	~0.11	~28
4	7.85~8.06	257~359	0.11~0.15	28~55
3	7.75~7.85	359~423	0.15~0.17	55~140
2	7.47~7.75	423~455	0.17~0.21	140~188
1	~7.47	455~	0.21~	188~

Along the same way, we chose 5 discrete regions of delay, jitter and loss rate (see Table 1). Among the 5 classes of parameter levels, we used only 4 levels in simulation because any level-1 value of each parameter always generated intolerable video quality.

We then have in total 4(parameter) ^ 4(level) = 256 combination sets (b, d, j, l) to measure the effects of each parameter sets to the quality of service at the moment.

C. Client PC

At the client PC, testers evaluated the streaming video quality. In the simulation, we used a simple binary quality measurement that only evaluates the quality as “Good” or “Bad” [10] in order to find out the percentile of dissatisfaction of users, or “unacceptability rate”. For example, if one tester out of the 16 testers notified “Bad”, the unacceptability rate is 1/16= 6.25% at the moment. If two people showed “Bad”, then the unacceptability rate becomes 2/16=12.5%

IV. SIMULATION RESULTS

Figures 2-5 show the simulation result, where X axis represents the unacceptability rate evaluated by the 16 testers. Y axis denotes the probability of occurrence of network parameters (b, d, j, l) for different quality levels; i.e., b5 is a typical value of bandwidth for Best (e.g., over 8.06Mbps), b4 for Good (e.g. 7.85~8.06Mbps), b3 for Moderate, and b2 for Bad.

For example, in Figure 2, at unacceptability rate 6.25% (at the most left-hand side), b5 occurred 3 times out of four tests (3/4=0.75), and b4 occurred once (1/4=0.25). Y axis denotes the probability of occurrence of bandwidth levels, and X axis denotes the unacceptability rate, where TL (Trend Line) is approximation plot of the dots. Here unacceptability rate 6.25% means a good quality because only one tester out of 16 felt Bad quality, and 15 felt Good. When the video is in good quality (i.e., when low unacceptability rate in X axis), there is no low level of bandwidth (b2) cases (in Figure 2, the “x” marks the presence of b2). As long as the video quality is decreasing (i.e., unacceptability rate increase in the X axis), we can find more occurrence of b2 (see the trend line “TL-b2” in Figure 2 is increasing along with the X axis). When the unacceptability rate increases, we can find low occurrence of b5 (see the trend line ‘TL-b5” is decreasing in Figure 2). This represents that video quality is strongly dependent on the bandwidth levels in the various combination set of (b, d, j, l).

Figure 3 shows the dependency of delay levels (d5~d2) on the video quality. Y axis denotes the probability of delay and X axis denotes the unacceptability rate. In Figure 3, we can find strong dependency of delay on the video quality because as the unacceptability rate increase, the TL-d5 de-

 ISBN: 978-1-61208-242-4

creases down and the TL-d2 increase up sharply. We can find Figure 3 shows similar pattern to Figure 2, which means that delay affects the video quality in a similar manner like bandwidth. It is noted that the levels (b5~b2) or (d5~d2) were determined from a level normalization process as summarized in table 1.

Dependency of jitter and loss on the video quality are shown in Figures 4 and 5, respectively. From Figures 4 and 5, it is shown that jitter and loss did not affect much on the video quality. It can be said that jitter is less sensitive to the quality of video comparing to the bandwidth or delay.

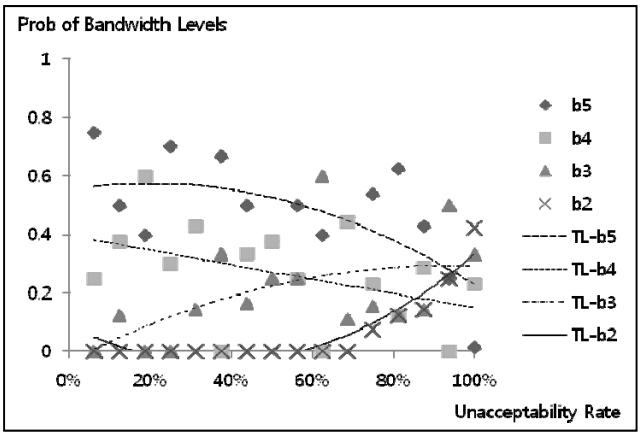


Figure 2. Dependency of bandwidth levels (b5~b2) on the video quality. Y axis denotes the probability of bandwidth levels, and X axis denotes the unacceptability rate, where TL (Trend Line) is approximation plot of the dots.

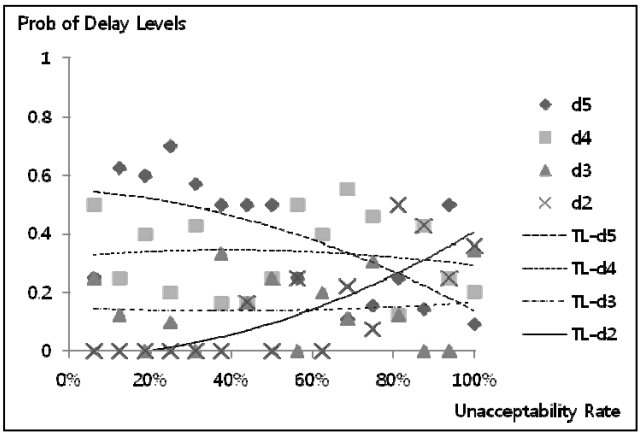


Figure 3. Dependency of delay levels (d5~d2) on the video quality. Y axis denotes the probability of delay levels, and X axis denotes the unacceptability rate, where TL (Trend Line) is approximation plot of the dots.

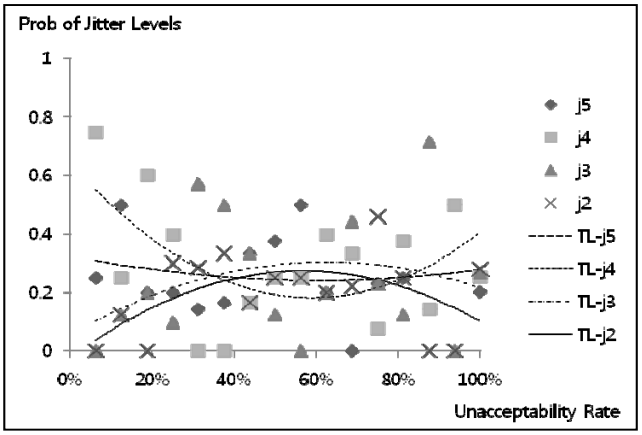


Figure 4. Dependency of jitter levels (j5~j2) on the video quality. Y axis denotes the probability of jitter levels, and X axis denotes the unacceptability rate, where TL (Trend Line) is approximation plot of the dots.

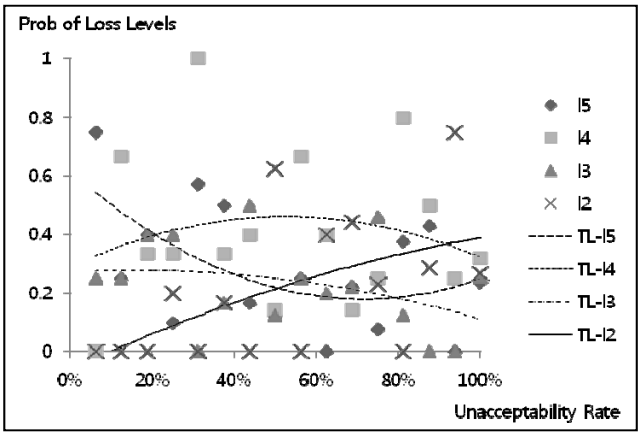


Figure 5. Dependency of loss levels (l5~l2) on the video quality. Y axis denotes the probability of loss levels, and X axis denotes the unacceptability rate, where TL (Trend Line) is approximation plot of the dots.

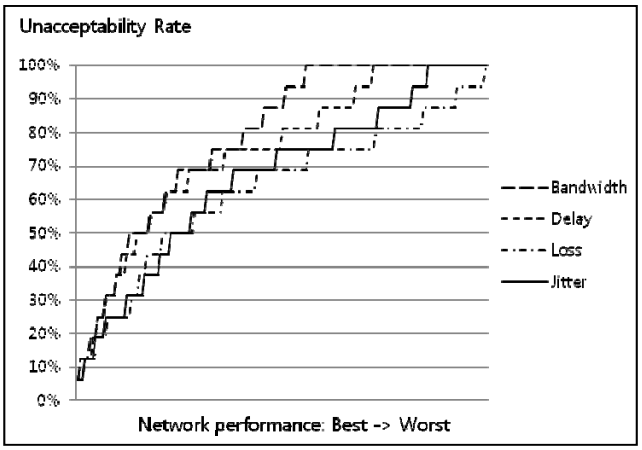


Figure 6. Comparison of the sensitivity of (b, d, j, l) to the quality variation (unacceptability rates).

 ISBN: 978-1-61208-242-4

Figure 6 compares the sensitivity of (b, d, j, l) to the quality variation (unacceptability rates). In Figure 6, the X axis shows the number of experiment cases under the assumption of that the distribution of each level of parameter is even. In other words, under the same distribution of parameter values, we can see that as the video quality is decreasing (going right in the X axis) the bandwidth gives more influence (i.e., sensitive) to the quality comparing to other parameters. The next sensitive parameter is delay and the next one is jitter.

V. CONCLUSION AND FURTHER WORKS

In the paper, we compared the influence of network parameters to the quality of streaming video. In order to perform the simulation within a finite number of tests, and compare them with evenly distributed patterns, we used a discrete set of levels for each parameter: bandwidth, delay, jitter and loss. Level 5 is Best quality, and level 4 is Good, level 3 is Moderate, and Level 2 is Bad. We found that bandwidth and delay affected directly the quality of video rather than the jitter or loss rate. We compared the dependency and sensitivity of the parameters on the service quality.

Even though the simulation was performed by 16 testers, and only 256 combination of parameter set are used in this paper, a larger dataset from real communication network service will provide a more accurate analysis.

ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0301-12-1004)

REFERENCES

[1] P. Wang, Y. Yemini, D. Florissi, P. Florissi, and J. Zinky, “Experimental QoS Performances of Multimedia Applications”, INFOCOM 2000, March, 2000

[2] Tao Yu; Baoyao Zhou; Qinghu Li; Rui Liu; Weihong Wang; Cheng Chang, “The service architecture of real-time video analytic system”, Service-Oriented Computing and Applications (SOCA), 2009, pp. 1-8.

[3] Sewook Oh, Seong Rae Park, “Providing qos for streaming traffic in mobile network with mobile router”, 3rd ACM workshop on QoS and security for wireless and mobile networks (Q2SWinet '07), 2007.

[4] Kostas E. Psannis, Yutaka Ishibashi, Marios G. Hadjinicolaou, “Qos for wireless interactive multimedia streaming”, 3rd ACM workshop on QoS and security for wireless and mobile networks (Q2SWinet '07), 2007.

[5] ITU-T Recommendation J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference”, 2008.

[6] Cisco, Internetworking Technology Handbook, “chap 49: Quality of Service Networking”.

[7] K. Piamrat, C. Viho, J-M. Bonnin, A. Ksentini, “Quality of Experience Measurements for Video Streaming over Wireless Networks”, Information Technology: New Generations, Sixth International Conference (ITNG ’09), 2009, pp. 1184 – 1189.

[8] Liuming Lu, Xiaoyuan Lu, Jin Li, “Quality Monitoring of Streaming Video Based on Packet-loss Artifacts”, Wireless and Optical Communications Conference (WOCC), 2010.

[9] http://www.ticomsoft.com/products/shunra/

[10] HwaJong Kim, KyoungHyoun Lee, Jie Zhang, “In-service Feedback QoE Framework”, Communication Theory, Reliability, and Quality of Service (CTRQ), 2010, pp. 135-138.

 ISBN: 978-1-61208-242-4

Modeling Team-Compatibility Factors Using a Semi-Markov Decision Process: A Data-Driven Framework for Performance Analysis in Soccer

Ali Jarvandi
Department of Engineering Management and Systems Engineering
George Washington University
Washington, DC USA
ajarvandi@gmail.com

Thomas Mazzuchi, Shahram Sarkani
Department of Engineering Management and Systems Engineering
George Washington University
Washington, DC USA
emseocp@gwu.edu

***Abstract*—Player selection is one of the great challenges of professional soccer clubs. Despite extensive use of performance data, a large number of player transfers at the highest level of club soccer have less than satisfactory outcome. This paper proposes a Semi-Markov Decision Process framework to model the dependencies between players in a team and its effects on individual and team performance. The study uses data from the English Premier League to artificially replace players in their prospective teams and approximate the outcome using a simulation. A comparison between the expected and actual performance determines the goodness of the model. In early experiments, the output of the model has correctly identified the trend changes in scoring and conceding goals and provided a high correlation with actual performance.**

Keywords-Markov Modeling; Decision Process; Sports; Simulation.

I. INTRODUCTION

Player transfers are a significant part of each soccer club's plan of action towards technical and financial success. This process has been traditionally guided by observation-based scouting. With the significant progress in data collection tools and methods in the recent years, professional soccer clubs have been collecting large amounts of data on their prospective players in order to make better transfer decisions. However, a preliminary study shows that about 40% of major signings by top European clubs between years 2010 and 2011 have been unsuccessful [1]. This translates to millions of dollars of loss for clubs. While detailed data is available on different aspects of players' performance, the challenge is to accurately utilize large amounts of data on various parameters towards making a single decision as to whether or not hire a player. This problem is particularly difficult due to the highly stochastic nature of the game as well as the interdependence between players, known as the compatibility factor [2]. While the challenge remains to develop an end-to-end model that addresses the dynamics of the game fairly accurately, there has been much progress in developing statistical models describing individual aspects of the game such as formation [3], possession [4], and goal scoring [5]. Though these models have offered great value to clubs in terms of making tactical decisions and designing training sessions, they have not been as useful in player selection. This is primarily because these models have not been applied in a context that addresses the two important issues of stochastic behavior and team compatibility.

This study is proposing a data-driven framework that incorporates existing sub-models in an end-to-end simulation in order to approximate player and team performance. Using a Semi-Markov Decision Process, this approach provides a stochastic foundation that reflects the nature of the game. In addition, the proposed approach offers the capability to artificially replace players in various teams and therefore include the effect of team compatibility in the estimated team performance. The historic performance data on players' technical and decision making attributes are used to construct the decision likelihoods and Transition Probability Matrices (TPM) needed for a Semi-Markov Decision Process. Finally, a strategy is developed to both test model accuracy and translate the output into specific predictions within the obtained accuracy. The ultimate goal of this study is to provide a decision support tool that assists decision makers in achieving a higher success rate in transfer decisions.

This paper provides an overview of the problem, state of the art, proposed approach, modeling strategies, and accuracy measures. Finally, the preliminary results have been listed and potential future directions have been discussed.

II. STATE OF THE ART

The studies of team selection in sports have been highly influenced by deterministic operations research. Boon and Sierksma modeled the team selection process using an integer programming approach in 2003 [6]. Also Dobson and Goddard computed the expected payoff associated with each playing strategy in 2010 [7]. In 2011, William A. Young developed a Knapsack model to approximate the compatibility between players in American Football using players' key individual attributes. While all of these studies provide valuable insight into the expected team performance, none of them has utilized players' decision making data or attempted to model the flow of the game.

This study considers the effects of individual decisions and interactions and offers the following advantages over the other methods:

 ISBN: 978-1-61208-242-4

A. *Using each player's decision making attributes in addition to technical attributes to model the flow of the game*

B. *Applying different sub-models to model various parts of the game with higher accuracy*

C. *Utilizing data in a stochastic framework, which is more reflective of the game*

D. *The ability to provide insight into the details of team and player performane as a result of modeling the entire flow of the system in approximately 1500 iterations per game*

III. METHODOLOGY

This study utilizes player performance data to model the interactions between players in a network. This leads to an approximation of the context generated for each player, which is then used to compute expected contribution to team performance in a simulation. The model is using a discrete time Semi-Markov Decision Process that provides players with a set of available decisions in each state at time t. Then the probability of transitioning to each state at time t+1 is computed by the probability of taking a decision multiplied by the corresponding value in the TPM. For example, a player in state one has four available decisions consisting of short pass, long pass, shoot, and dribble. The player's historical data provides the percentage of the time that each of these decisions is made and the success rate associated with each decision. This feeds the likelihood fields and the TPM needed for that player in state one. Fig. 1 shows the transition for a given decision at time t. In the figure, $P(D_1, S_k)$ represents the probability of transitioning to state k under decision 1.

The possible states for each player at a given instant in the game have been defined as following:

A. *Player has possession of the ball*

B. *A teammate has possession of the ball*

C. *The opposition has possession of the ball*

D. *Neither team has full possession of the ball*

Each of these states is associated with a set of available decisions that is mutually exclusive from the decisions available in other states. This results in 4 TPMs for each player, corresponding to each possible state. Table I shows a sample TPM from the model.

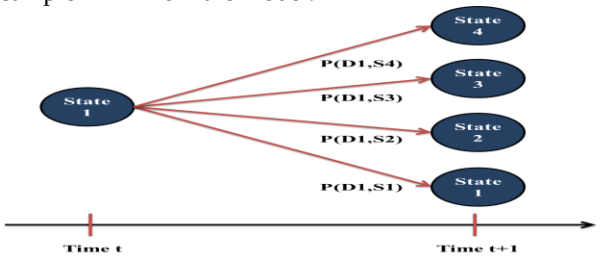


Figure 1. Transition from state 1 under decision 1.

TABLE I. TPM AND DECISION LIKELIHOOD IN STATE 1

Likelihood	Cesc Fabregas	State			
	Decision	1	2	3	4
0.8564	Short Pass	0	0.8241	0.1759	0
0.0765	long Pass	0	0.5954	0.4046	0
0.0414	Shoot	0	0	0.7042	0.2958
0.0257	Dribble	0.55	0	0.45	0

In addition to the transition probability matrices, which reflect players' individual trends, this study uses dependency matrices for various decisions to capture the effects of individual decisions on team trends. For instance, an 11x11 matrix represents the probability of each prayer playing a short pass to each teammate, given that a decision has been made to play a short pass.

Finally, two logical flows are developed to capture the processes for scoring and conceding goals. For goal scoring, two attributes of chance creation and conversion are captured for each player. These strictly individual attributes in the context of Transition and Dependency Matrices previously formed, will determine the team's expected scoring record, goal scorers, and more frequent plays. The process of conceding goals however, cannot be modeled in a similar way due to the lack of contribution measures for individual players. Therefore, it is significantly more relevant to model the process of conceding goals as a team process rather than sum of individual contributions. Previous studies on goal scoring in soccer have suggested different probability models based on two parameters: the position in which the scoring team gains possession of the ball, and the number of passes that are exchanged among the scoring team that lead to a goal. Using these attributes in a reverse way can generate a model for conceding goals. As a result, each possession for the opponent carries two probabilities that contribute to the chance of conceding a goal: Starting Probability, which is defined, based on the position where possession of the ball has been lost; and Carrying Probability, which is a function of the number of passes exchanged within the opponent. A model containing these two probabilities can lead to an accurate estimation of conceding goals in different conditions.

IV. MODEL OUTPUT

A regression analysis of the goal differential and number of points obtained in three of the top European leagues (Spain, England, and Italy) shows that the number of points obtained by each team can be predicted with a large degree of confidence (r^2>0.9 in each of the three leagues) based on goal differential by using historical data (Fig. 2).

 ISBN: 978-1-61208-242-4

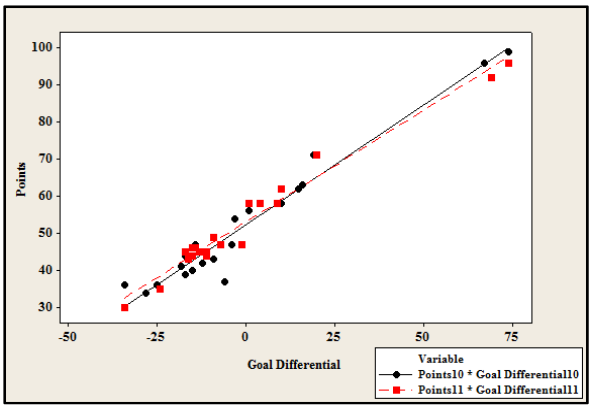


Figure 2. Points versus goal differential - Spain, 09/10 and 10/11 seasons.

The high correlation between points and goal differential in consecutive years enables the model to use expected number of goals scored and allowed as the ultimate team performance measure. Therefore, the significance of each player's technical and decision-making attributes is measured by its estimated contribution to scoring or conceding goals. The final model output will then be the expected number of goals scored and conceded in a large number of games. The difference between the expected changes in the number of scored and conceded goals and the team's previous record determines the estimated contribution of the new players to the team performance.

V. DATA AND RESULTS

The data used in this study includes game-by-game and season-by-season performance data for all players in the English Premier League between 2008/09 and 2011/12 seasons. The data also includes each player's overall playing minutes in the season, which is used to normalize player performance data with respect to other players. Also, while accounting for injuries is not in the initial scope of this research, it is important to note that in some cases raw data without normalization for injury periods can provide insight into the impacts of players' injury prospect on their overall performance. To analyze each transfer from team A to team B, three sets of data will be used: 1) Data from team A in year n, 2) Data from team B in year n, 3) Data from team B in year n+1. Using this criteria, 116 transfers have been identified that can be analyzed using the current dataset.

To approximate the expected team performance, the data for the player of interest from team A is replaced in the same playing position in team B and the player's impact will be measured by the difference in the expected number of goals scored and conceded compared to the output of the model for the original dataset 2. Finally, a comparison between the expected change in goal differential and the difference between the normalized goal differential from team B in year n+1 with the player of interest on and off the pitch determines the model accuracy. In the rare cases when a player plays the entire season, model accuracy is simply measured by a comparison between the expected and actual change in goal differential. For each transfer, model accuracy is measured using two parameters: Trend Identification and Correlation. Trend Identification is the measure for the success rate of predictions made by the model and refers to the overall effect of each transfer on team performance. The values for this attribute can be "Positive", "Negative", or "Insignificant". Assuming that the model correctly predicts this value, Correlation is computed. This measure determines the accuracy of the model in approximating the magnitude of the difference in the expected performance caused by the introduction of the new player. Currently, the model has been run for 5 transfers and has provided 100% success in trend identification and 81.3% average correlation with the actual performance. While these numbers can change with a larger dataset, the obtained accuracy is significantly higher than the one in the current transfer market.

VI. CONCLUSION AND FUTURE WORK

The proposed study provides a framework for utilizing player performance data for approximating the expected performance of a prospective player within a given context. As a decision support tool, this model will assist clubs in increasing their success rate in player transfers and reaching higher efficiency in budget allocation. Potential future work on this topic include a risk analysis of squad selection with this method with respect to potential injuries, the effect of substitute players on model accuracy, and the study of transfers from a league to another.

ACKNOWLEDGMENT

The authors would like to thank Dr. David F. Rico for his continuous support in proposing the original idea.

REFERENCES

[1] J. Ali, "A System Approach to Team Building in Soccer", Unpublished.

[2] W.A. Young, "A Team-Compatibility Decision Support System to Model the NFL Knapsack Problem: An Introduction to HEART," unpublished.

[3] N. Hirotsu and M. Wright, "Determining the best strategy for changing the configuration of a football team," Journal of the Operational Research Society, vol. 54, pp. 878-887, 2003.

[4] M. Shafizadeh, S. Gray, J. Sproule, and T. McMorris, "An exploratory analysis of losing possession in professional soccer," International Journal of Performance Analysis in Sport, vol. 12, pp. 14-23, 2012.

[5] A. Tenga and E. Sigmundstad, "Characteristics of goal-scoring possessions in open play: Comparing the top, in-between and bottom teams from professional soccer league," International Journal of Performance Analysis in Sport, vol. 11, pp. 545-552, 2011.

[6] B. Boon and G. Sierksma, "Team formation: Matching quality supply and quality demand," European Journal of Operational Research, vol. 148, pp. 277-292, 2003.

[7] S. Dobson and J. Goddard, "Optimizing strategic behavior in a dynamic setting in professional team sports," European Journal of Operational Research, vol. 205, pp.661-669, 2010.

[8] P. O'Donoghue, K. Papadimitriou, V. Gourgoulis, and K. Haralambis, "Statistical methods in performance analysis: an example from international soccer," International Journal of Performance Analysis in Sport, vol. 12, pp. 144-155, 2012.

 ISBN: 978-1-61208-242-4

Design and Operation of the Electronic Record Systems of the US Courts are Linked to Failing Banking Regulation

Joseph Zernik
Human Rights Alert (NGO)
Jerusalem
e-mail: 123456xyz@gmail.com

***Abstract*— The current report is based on data mining of the information systems of the US courts – PACER (Public Access to Court Electronic Records) and CM/ECF (Case Management/Electronic Court Filing) in general, and review of the electronic records in a landmark litigation under the current financial crisis: *Securities and Exchange Commission v Bank of America Corporation* (2009-10). The case originated in the unlawful taking of $5.8 billion by banking executives, and concluded with the executives never returning the funds to the stockholders and with no individual being held accountable. The case was covered numerous times by major US and world media. Data mining of records of the US courts from coast to coast reveals built-in deficiencies in validity and integrity of the PACER and CM/ECF. The case study documents missing and invalid litigation records, leading to the conclusion that the case as a whole amounts to simulated litigation. A number of corrective measures are outlined, including publicly and legally accountable functional logic verification of PACER and CM/ECF, and correction of the defective signature and authentication procedures now implemented in the systems. This study highlights the significance of application of data mining to the target area of court records in particular, and records of the justice system in general. It is also a call for action by computing experts and data mining experts in for the safeguard of Human Rights and integrity of governments in the Digital Era.**

Keywords- e-Courts; e-Government; United States; Banking Regulation; PACER; CM/ECF.

I. INTRODUCTION

The US government has completed a decade-long project of transition to electronic administration of the US courts at a cost that is estimated at several billion US dollars. Courts of nations around the world are going through similar processes, and United Nations reports on judicial integrity promote such transition. Obviously, valid electronic record systems could have enhanced the integrity and transparency of the courts.

Record keeping under paper administration of the courts has evolved over centuries and formed the core of due process and fair hearing. The transition to electronic administration of the courts affected a sea change in such court procedures.

Two systems were implemented in the US district courts and US courts of appeals:

- PACER – for Public Access to Court Electronic Records, and
- CM/ECF – for Case Management and Electronic Court Filing.

Previous reports inspected the validity and integrity of the systems, and identified core defects inherent in their design and operation. [1,2]

Fraud in the state and US courts has been increasingly recognized as a key part of the current financial crisis. [3-6] A legal expert opined, "… it's difficult to find a fraud of this size on the U.S. court system in U.S. history… I can't think of one where you have literally tens of thousands of fraudulent documents filed in tens of thousands of cases." [7] Concern has also been repeatedly voiced with the ineffectiveness of US banking regulation. [7]

Other papers have documented fraud in the litigation of cases in the US district courts, in the US court of appeals, and the US Supreme Court. [8-10]

Under such circumstances, this study inspects the information systems, implemented by the US courts, and their implications in the litigation of a landmark case under the current financial crisis: *Securities and Exchange Commission v Bank of America Corporation* (1:09-cv-06829). In this case, the US Securities and Exchange Commission (SEC) purportedly prosecuted Bank of America Corporation (BAC) in the US District Court, Southern District of New York, for violations of securities laws, related to the unlawful taking by executives of $5.8 billion. Financial analysts described the underlying matter as a 'criminal conspiracy' and 'bigger than Watergate?'

Based on data mining in this unique target area, this study claims that instead of enhancing integrity and transparency of the courts, the electronic record systems have enabled unprecedented, widespread corruption in the US courts.

This study highlights the significance of application of data mining techniques to the target area of court records. It is also a call for action by computing experts in general and data mining experts in particular for the safeguard of Human Rights and integrity of governments in the Digital Era.

 ISBN: 978-1-61208-242-4

A. Authorities of Judges and Clerks

Generally, the courts are described as consisting of two arms: the judicial arm and the ministerial (clerical) arm. Both the judges and the clerks must be duly appointed and hold their positions under Oath of Office. A judge is authorized to decide on a matter, duly presented to him by parties in a given matter. The clerk is the legal custodian of the data base of court records.

Therefore, a judge may decide and issue a signed decision in a given matter. However, such record is not a valid court record, until it is duly inspected for validity (case number, case caption, signatures, assignment of the judge to the specific case, etc) and formally entered into the data base under the signature of a named, authorized clerk.

The entry of a judicial record into the data base is inseparable from authentication of the record through its service on all parties in the case. Accordingly, for example, the authentication record, accompanying each and every judicial record and signed by the Clerk in the Superior Court of California, is titled "Certificate of Mailing and Notice of Entry by the Clerk."

B. PACER and CM/ECF

Critical deficiencies were previously identified in design and operation of PACER and CM/ECF: [10]

- Both the courts and the US Congress failed to establish by law the new electronic court procedures, inherent in the systems. Particularly, no law today defines the form of valid and publicly recognizable digital signatures of judges and clerks. (Figures 1,2, and 3; see Online Appendix 1)
- Public access to the electronic authentication records, which are maintained in CM/ECF, but not in PACER, is routinely denied. The corresponding records were essential part of the public records under paper administration of the courts.
- The systems enable attorneys to appear, file, and purportedly enter records with no prior review by the clerks of the courts, and with no power of attorney by their purported clients. [11,12] The systems also enable unauthorized court personnel to issue and publish online court records – dockets, minutes, orders, and judgments, which were never authenticated by the duly authorized clerks. [10]
- There is no evidence that the systems were subjected to functional logic verification in a publicly and legally accountable manner.

The outcome of such conditions, as demonstrated in this study is that the public is unable to distinguish between valid and void court records. With it, the systems enable the courts to conduct of simulated litigation, otherwise known as "Fraud on the Court". [13]

II. METHODS

A. General Approach to data mining of the records of the US courts

The research presented in this paper was narrowly focused on analysis of integrity of the electronic record systems in the national courts (US district courts). The study was not based on legal analysis of the records, or challenges to the rationale of the adjudication in specific decisions, except for the laws pertaining to the maintenance of court records.

Rules, pertaining to the operation of the systems were investigated through review of the Users' Manuals of the various courts, where available. Such manual provided limited rules at best.

Therefore, rules of the systems had to be inferred through data mining.

UNITED STATES DISTRICT COURT
FOR THE
DISTRICT OF <name of district>

Case Number: <number>
<name1>, *Plaintiff*,
v.
<name2>, *Defendant*.

CERTIFICATE OF SERVICE

I, the undersigned, hereby certify that I am an employee in the Office of the Clerk, U.S. District Court, Northern District of California.

That on <date>, I SERVED a true and correct copy(ies) of the attached, by placing said copy(ies) in a postage paid envelope addressed to the person(s) hereinafter listed, by depositing said envelope in the U.S. Mail, or by placing said copy(ies) into an inter-office delivery receptacle located in the Clerk's office.

<service list>

Dated: <date>
<name of individual>, Clerk
[wet graphic signature]

_____
By: <name of individual>, Deputy Clerk

Figure 1. Paper-based Certificate of Service, as Used by the Clerks of the US District Courts Prior to Implementation of PACER and CM/ECF, the Electronic Record Systems of the US Courts.

The Certificate of Service was titled "Certificate", it included a certification statement, "I certify…", the name and authority of the individual executing the certification, on behalf of the clerk of the court, and his/her hand signature. The certification records were essential part of the public records in the paper court files. None of these elements were preserved in PACER and CM/ECF (Figure 2; see Online Appendix 1).

Subsequently, irregularities, or contradictory data in date, signature, certification, and registration procedures were examined through data mining methods, executed on the online public records of the courts, through various search options, provided in the systems themselves.

Initially, integrity of the basic components of the systems were examined: indices of all cases, calendars, dockets (lists of records in a given file), indices of decisions, and compliance of these components with the *Federal Rules of*

 ISBN: 978-1-61208-242-4

Civil Procedure, pertaining to the maintenance of court records and the duties of the clerks in this regard.
The analysis was also based on consultations with Israeli law computing/cryptology experts. No US law or computing expert, who was approached, agreed to discuss the matter.

B. *Information regarding PACER and CM/ECF in the US District Court, Southern District of New York*

The platforms of PACER and CM/ECF in the US District Court, Southern District of New York, are the same as previously described in other US district courts. [2, 10] The electronic authentication records (NEFs – Notices of Electronic Filings) are excluded from public access in PACER. The NEFs are today accessible only through CM/ECF, only by court personnel and by attorneys, who are authorized by the courts in particular cases.

The *Local Rules* of the US District Court, Southern District of New York, as downloaded from the Court's website, were reviewed, particularly relative to the following matters: [14]

- Operations of PACER and CM/ECF and court procedures inherent in them;
- Certification of court records under electronic filing;
- Issuance, docketing, and execution of the service of summons and complaint;
- Certification of attorneys as Attorneys of Record upon appearance, and
- Definition of valid digital signatures under electronic filing.

As was the case in other US courts, which were examined, the *Local Rules* of the US District Court, Southern District of New York, failed to provide clear and unambiguous details of electronic filing and electronic court records procedures in general, and in particular relative to the matters listed above.

Further information regarding the design and operations of PACER and CM/ECF was found in the *Electronic Case Filing Rules & Instructions* and the *CM / ECF Attorney Civil Dictionary* of the US District Court, Southern District of New York. [14] However, such sources are not valid legal sources for establishing the validity of court records.

Users' manuals, albeit not valid court records, provided additional information regarding the design and operation of PACER and CM/ECF in the various US district courts, which were inspected. [15] No public access was found online to the Users' Manual of the US District Court, Southern District of New York.

C. *Court Records in Litigation of SEC v BAC*

The primary source of records was through online public access to PACER records.

Additionally, efforts were made to obtain missing records, where access in PACER was not available, through requests, which were forwarded to the office of the Clerk of the Court, to attorneys for SEC and BAC, to the office of the Chair of SEC, and to the office of BAC President. *Freedom of Information Act* (FOIA) requests were filed on the office of SEC and the office of Fair Fund Administrator (detailed below).

D. *Media Reports*

Major media outlets: New York Times, Washington Post, Wall Street Journal, and Times of London were searched online for media coverage of the litigation.

III. RESULTS

A. *Public Access to Court Records through PACER*

Public access to US court records is effetively permitted only through PACER. However, the identity of the server, from which the records are displayed and downloaded is unverified. A typical browser "page info" statement for PACER is:

Website Identity
Website: **http://www.pacer.gov/**
Ownership: **This website does not supply ownership information.**
Verified by: **Not specified**
Technical Details
Connection Not Encrypted.

B. *Public Access to the litigation records in SEC v BAC*

1) *NEFs*

As is the case in all other US district courts, which were inspected, public access to the authentication records (NEFs) of the US District Court, Southern District of New York in PACER is denied.

Although never stated in the *Rules of Court* of the US District Court, Southern District of New York, the evidence shows that, as is the case in other US courts, the NEFs are in fact deemed the authentication records of the Court. No other form of authentication record, pertaining to judicial records (minutes, orders, judgments) was found in the PACER docket.

The form of the NEFs of the US District Court, Southern District of New York, is identical to the NEFs, as discovered in other US courts. The NEF, in and of itself, should be considered invalid Certificate of Service. It includes no mention of certification in its title, it includes no certification statement - "I certify…", it never invokes the authority of the Clerk of the Court, it never names the person, who issued the NEF, and it includes neither a graphic hand-signature, nor a valid digital signature of an authorized individual, affixed as a symbol of intent to take responsibility. Instead, it only includes a machine generated, encrypted checksum string. (Figures 1, 2; see Online Appendix 1)

2) *Calendar and Docket Activity Report*

Furthermore, the US District Court, Southern District of New York, unreasonably limits public access to the Calendars of the Courts, in contrast with other US district courts. Access is permitted only to the current seven days window. Therefore, conduct of the "off the record"

 ISBN: 978-1-61208-242-4

proceedings and other proceedings with no valid minutes cannot be confirmed. Likewise, the US District Court, Southern District of New York, denies access to the Docket Activity Report – critical data for confirmation of docket notations.

3) Summons, Minutes

Although the *Federal Rules of Civil Procedure* prescribe the docketing of the summons as issued by the Clerk of the Court and the minutes of the proceedings, these records were missing from PACER docket of the case at hand.

C. Access to Litigation Records from Other Sources

Requests for copies of the missing litigation records, which were forwarded to the office of the Clerk of the Court, to attorneys for parties in the case, to the office of SEC Chair, and to the office of President of BAC, were all refused.

Eventually, only one NEF was obtained from the office of the Fair Fund Administrator (2; see Online Appendix 1). Access to the other NEFs was denied by the same office with no explanation at all.

The summons, as issued by the Clerk of the Court, was eventually obtained in response to a *Freedom of Information Act* (FOIA) request on SEC (Figure 4; see Online Appendix 1). [16] However, the same office failed to produce the other missing records of the NEFs in this case under the claim that "these documents are not agency records under the FOIA." Such claim contradicted the response of the same office on the summons in this case. Conduct of SEC in this matter amounts to selective compliance with the *Freedom of Information Act*.

The September 9, 2010 FOIA response letter from SEC states:

> This letter is a final response to your request, dated January 26, 2010, and received in this office on February 2, 2010, for certain information concerning SEC v. Bank of America Corporation (BAC)(1:09-cv-06829), United States District Court for the Southern District of New York. Specifically, you request copies of the following:
>
> a) Summons, as issued by clerk under such caption; NEF (Notice of Electronic Filing) pertaining to the Summons, as issued by clerk under such caption;
> b) Summons, as executed under such caption.
> c) NEF (Notice of Electronic Filing) pertaining to the Summons, as executed under such caption.
> d) NEF (Notice of Electronic Filing) pertaining to the complaint under such caption.
> e) NEF (Notice of Electronic Filings) for each and every court order under such caption.
>
> In response to a) and c) above please find attached a copy of the summons issued in SEC v. Bank of America Corporation (BAC)(1:09-cv-06829).
>
> With respect to the remainder of your request related to Notices of Electronic Filing generated by the United States District Court for the Southern District of New York, please be advised that these documents are not agency records under the FOIA. Consequently, we are considering your request closed.

D. PACER Chronology of SEC v BAC (1:09-cv-06829)

1) The PACER Docket as a Whole

A total of 19 purported proceedings are listed in the PACER docket (see the full PACER docket and linked records under the Online Appendix). For two of the proceedings, no docket entries are found at all. These two proceedings should be deemed 'off the record' proceedings. For the remaining 17 proceedings, invalid docket entries are found - no minutes records at all are linked to the docket listings, no docket numbers are designated, and no information regarding content of the proceedings is provided in the docketing text.

Moreover, based on the design of the NEFs (see below), it is impossible that such docket entries were authenticated. Therefore, these are invalid docket entries, with no corresponding valid court records. (Figure 3; see Online Appendix 1) These remaining 17 proceedings should be deemed simulated court proceedings.

Of such 17 minutes with no records, 16 are listed as 'entered' by an individual only identified as 'mro'. The full name and authority of the person remain unknown. In contrast, out of a total of 24 docket listings of orders and the judgment, where docket number is designated, where a record is linked, and where informative docketing text is provided, not a single item is entered by 'mro'. The Docketing Department of the Court confirmed that docketing in the case was not performed by authorized Deputy Clerks, but refused to disclose the names of the individuals involved.

Entry of transactions in the PACER dockets by unauthorized court personnel, unbound by Oath of Office, was documented in other US district courts as well. Beyond undermining the integrity of the dockets, such practices demonstrate the lack of security and validity of PACER and CM/ECF as a whole.

Overall, nowhere in the PACER docket is the authority of the office of the Clerk of the US Court invoked, neither is the name, nor the authority of any individual as Deputy Clerk to be found.

2) Individual PACER Docket Records

A detailed review of the entries in the PACER docket, and their validity, or lack thereof, is provided in the Online Appendix. Here, only highlights are provided.

a. August 3, 2009 – Summons issued

The PACER docket in this case states that on August 3, 2009, summons was issued as to BAC. However, absent a Docket Number, the text cannot possibly be deemed a valid docket entry. Additionally, no link is provided in the PACER docket to the summons record itself. Therefore, no

 ISBN: 978-1-61208-242-4

electronic authentication could have been issued through an NEF under such circumstances on the summons.

Access to the summons as issued and as served was repeatedly requested from the office of the Clerk of the Court. Access was denied.

Eventually, a copy of the summons, as issued by the Clerk of the Court, was obtained through a *Freedom of Information Act* (FOIA) response from SEC. (Figure 4; see Online Appendix 1) The summons as issued by the Clerk of the Court in this case is unsigned, and bears no seal of the Court. Therefore, the summons is in fact a simulated summons record. [13] The failure to issue and docket valid summons is consistent with the intent to conduct simulated litigation from the start.

b. Service of Process

Nowhere in the docket is there any indication that service of the summons was in fact executed, alternatively, that service of the summons was waived.

In the FOIA response, the unsigned summons record was provided as both the record of the summons as issued and the summons as executed. (Figure 4; see Online Appendix 1) Needless to say, execution of service of an unsigned, fraudulent summons is invalid execution.

Therefore, the service of process in this case was simulated, service of process.

c. August 3, 2009 - Motion for entry of Settlement Agreement

Media reported from the onset of the litigation that a proposed Settlement Agreement for $33 million, to be paid by BAC to the US government, was filed on August 3, 2009. Later, several docket notations refer to the pending initial Settlement Agreement.

On September 24, 2009, Order was entered in the docket, purportedly rejecting the then pending proposed Settlement Agreement. However, no Motion for entry of the proposed Settlement Agreement and no proposed Settlement Agreement records are found in the docket.

The failure to file a Motion for Entry of the Settlement Agreement, and consequent issuance of an order, purported to deny the motion with no motion record are consistent with the conduct of simulated litigation.

d. August 10, 2009 – 'off the record' proceeding

There is no docket listing at all for the August 10, 2009 proceeding. However, evidence of its conduct is provided in the PACER docket entries #15,16, and 17.

The conduct of 'off the record' court proceeding is also consistent with the conduct of simulated litigation in this case.

e. August 25, 2009 - Order denying the initial proposed Settlement Agreement

The August 25, 2009 Order (Dkt #13) states:

> This Court has the obligation, within carefully prescribed limits, to determine whether the proposed Consent Judgment settling this case is fair, reasonable, adequate, and in the public interest.

The ruling on a matter that was never pending before the Court is also consistent with the conduct of simulated litigation in this case.

f. February 22, 2010, Final Consent Judgment

On February 22, 2010, the Final Consent Judgment (Dkt #97) was filed by Judge Jed Rakoff.

The docketing text of the Final Consent Judgment states:

> FINAL CONSENT JUDGMENT AS TO DEFENDANT BANK OF AMERICA CORPORATION # 10,0297 in favor of Securities and Exchange Commission against Bank of America Corporation in the amount of $ 150,000,001.00. (Signed by Judge Jed S. Rakoff on 2/24/2010) (Attachments: # 1 Notice of Right to Appeal) (dt) (Entered: 02/24/2010)

Access was repeatedly requested to the NEF of the Final Consent Judgment, absent which, the record cannot be deemed a valid court record. Access was denied in violation of the law (First Amendment).

3) Other PACER Records

Detailed review of the other PACER records is provided in the Online Appendix. Here only highlights are provided:

a. *August 3, 2009 - Civil Cover Sheet*

Records of the US District Court, Southern District of New York, prescribe that a Civil Cover Sheet be filed with the complaint. Likewise, the Civil Litigation Management Manual of the Judicial Council of the United States, [18] prescribes the filing of a Civil Cover Sheet as prerequisite for opening a new docket by the Clerk of the Court, no Civil Cover Sheet is found in the docket of the case at hand. In contrast, Civil Cover Sheets are routinely found in the PACER dockets with the complaints in other cases. As further noted in the *Rules of Court*, the Civil Cover Sheet must be part of Service of Process, together with the summons and complaint.

The failure to file and docket a Civil Cover Sheet is of particular significance, since it is one of the only records, where a publicly visible, hand signature of the Clerk of the Court, identified by name and authority, is still required.

The failure to file and serve a Civil Cover is consistent with the intent to conduct simulated litigation from the start.

4) Freedom of Information Act (FOIA) and Fair Fund Administrator Responses

While the Clerk of the Court refused to provide access to the missing court records in this case, which were and are public records by law, access was gained to certain missing records from other sources.

The FOIA response by SEC yielded a copy of a record that was described as "summons, as issued by clerk" and "summons, as executed". (Figure 4; see Online Appendix 1) However, the record produced in the FOIA response is an unsigned summons with no seal of the Court. Therefore, the record is a simulated summons.

The Fair Fund Administrator provided a copy of one NEF only (Figure 2; see Online Appendix 1). The same office refused to provide the NEFs of other records, most

 ISBN: 978-1-61208-242-4

notably the NEF of the Final Consent Judgment (Dkt #97). No explanation at all was provided by the Fair Fund Administrator for the refusal to provide copies of the additional NEFs.

5) Circumstances Surrounding the Litigation of SEC v BAC

The New York Times reported the news regarding the Final Consent Judgment in this case as follows: [17]

> In a ruling that freed Bank of America from some legal problems, a federal judge wrote on Monday that he had reluctantly approved a $150 million settlement with the Securities and Exchange Commission… "This court, while shaking its head, grants the S.E.C.'s motion and approves the proposed consent judgment," the judge wrote.

Independent investigation of events surrounding the BAC-Merrill Lynch merger by State of New York Attorney General Andrew Cuomo was summed up in his April 23, 2009 letter to the US Congress. [19] Financial analysts' responded the release of the letter and its attachments under headlines such as "Let the Criminal Indictments Begin: Paulson, Bernanke, Lewis", "Bigger Than Watergate?", and "Cuomo Unveils Paulson, Bernanke, Lewis Conspiracy". However, in contrast with such analysts' opinions, none of the perpetrators suffered any material consequences so far.

Judge Jed Rakoff, who presided in the case, is considered one of the most experienced and notable among US judges in matters pertaining to securities, white-collar crime, and racketeering.

Regarding Judge Jed Rakoff's conduct in this case, the *Wall Street Journal Law Blog* said: [20]

> Rakoff is currently proving himself to be, if nothing else, unafraid to single-handedly take on some heavyweight institutions and their lawyers.

Numerous attorneys appeared in the case, both for Plaintiff SEC - a government entity, and for Defendant BAC – a major financial institution. Over twenty Notices of Appearance appear in the PACER docket of the case.

In a September 22, 2009 report, the Washington Post quotes official statement by SEC:

> "[W]e will vigorously pursue our charges against Bank of America and take steps to prove our case in court," the SEC said in a statement. "We will use the additional discovery available in the litigation to further pursue the facts and determine whether to seek the court's permission to bring additional charges in this case."

It is practically impossible that the prominent attorneys, who appeared in the case for both SEC and BAC, were unaware of the simulated nature of the litigation.

The litigation of this case was extensively covered by major US and international media outlets, as a key litigation under the current global financial crisis. A total of ten (10) reports of the litigation were found in the New York Times, eight (8) reports in the Washington Post, and dozens of reports in the Wall Street Journal. Likewise, fourteen (14) reports related to the case were found in the Times of London. However, there is no reason to assume that any media, which reported on the litigation, ever tried, or gained access to records in the case beyond those accessible through the PACER docket, as described above. There also is no evidence that media ever reported that critical litigation records are missing. It is difficult to believe that experienced legal reporters of major media outlets never noticed the fatal flaws in the records and conduct of the litigation.

IV. DISCUSSION

A. Overall Conduct of the Litigation

The conduct of the US District Court in this matter should be reviewed in the context of the conduct of civil litigation as stipulated in the *Federal Rules of Civil Procedure*, and as outlined in the *Civil Litigation Management Manual, Second Edition* (2010) by the Judicial Conference of the United States. [18] Upon such review, a reasonable person would conclude that the litigation as a whole was never deemed by the US Court itself as valid litigation.

1) Issuance and Service of Simulated Summons

The summons is a critical record – it establishes the onset of litigation and also establishes the jurisdiction of a particular court in a particular matter on particular parties. Accordingly, the *Federal Rules of Civil Procedure* prescribe that the summons be docketed by the clerk of the court and therefore become a public record. Moreover, the US Code prescribes that the summons be issued under the signature of the Clerk of the Court and under the Seal of the Court.

The invalidity of the summons in this case undermines the validity of the litigation as a whole. However, the invalidity of the summons in this case could not be discerned in PACER, the public access system.

Since the issuance of the summons is the very first action in any case, the issuance of a simulated summons in this case also shows that the case was designed to be simulated litigation from the very start, through collusion of the Judge, the Clerk of the Court, and the attorneys for SEC and BAC. In fact, the case was intended from the start to be an inverse show trial. [21]

2) The Fairness Perspective

The outcome of the litigation should also be viewed from the fundamental fairness perspective: Individuals, who were banking executives, unlawfully took $5.8 billion from the shareholders of BAC.

In response, SEC and BAC proposed, concomitantly with the filing of the complaint, the initial Settlement Agreement, which would have imposed a fine of $33 milion on the stockholders - the victims of the unlawful conduct.

Eventually, the outcome of the litigation was that pursuant to the final Settlement Agreement, compensation in the sum of $150 million was paid by the shareholders to themselves. No funds were returned by any of the

 ISBN: 978-1-61208-242-4

perpetrators, and none of the perpetrators was held accountable for their unlawful conduct.

B. Human Errors, or Simulated litigation?

The conduct, documented in this study and in numerous other cases in the US district courts, the US courts of appeals, and the US Supreme Court cannot reasonably be deemed the outcome of human errors for the following reasons:

- Any valid and honest electronic record system of a courts must be secure enough, so that it would not enable unauthorized, unnamed persons, unbound by Oath of Office, to execute any transactions in the system.
- The authentication records, implemented in CM/ECF (the NEFs), bear neither the name, nor the authority and signature of the person, who issues the authentication record. Moreover, the NEFs are excluded from the data base of public records of the courts, and the courts deny access to these records in violation of First Amendment rights. There is no plausible explanation for such design of the systems and the universal denial of access to these records, which is consistent with integrity of the US courts.
- In all cases in the state and US courts, where simulated service of judicial records and their incorporation in the dockets was brought to the attention of the respective court, the courts refused to correct the dockets. Instead, the courts proceeded with the litigation of such cases, treating such simulated records as valid and effectual court records. [22,23] In doing so, the courts deliberately ignored the fact that the service of invalid judicial records and the publication of the same by the courts are the essence of the criminality, here referred to as "Simulating Legal Process", [13] and historically known as "Fraud on the Court". [24]

C. Transparency of the judicial process – public access to court records.

Court records are public records pursuant to the US law – the First Amendment. The landmark decision in this matter by the US Supreme Court in *Nixon v Warner Communications, 1978,* pertaining to the Nixon tapes, explicitly states that it only re-affirms existing law in this regard, and that public access to court records is essential, in order to enable to People "to keep a watchful eye on government" (including, but not limited to the judiciary).
As documented in this case and numerous others, today public access to court records in the United States is selective.

A major claim of this study is that the selective denial of public access to court records has been enabled in recent decades through the implementation of invalid electronic record systems in the courts. It effectively amounts to establishment of "double books" systems in the courts – in the US courts – PACER and CM/ECF. [25]

D. Missing Court Records – Cardinal Sign of Judicial Corruption.

United Nations reports on "Strengthening Judicial Integrity" advocate the implementation of electronic record systems in the courts, as a tool for enhancing the transparency and integrity of the courts. [29,30] However, the same United Nations reports list missing court records as a cardinal sign of judicial corruption.

In the case of PACER and CM/ECF, the implementation of invalid electronic record systems in the courts enables the conduct of litigation, where critical records are permanently missing, as demonstrated in the current study.

E. Banking Regulation

Conduct of the SEC, BAC, and the US District Court, Southern District of New York, as documented in the current study, stands contrary to ongoing efforts by the US Congress and repeated statements by the US Government, regarding efforts to restore the integrity of Banking Regulation in the United States and enhance the accountability of directors, executives, accountants, and attorneys, acting on behalf of corporations. In the wake of the Enron scandal, the US Congress passed the *Sarbanes-Oxley Act* (2002), in order to enhance the accountability of corporate officers, attorneys, and accountants. In the wake of the current financial crisis, the US Congress passed the *Fraud Enforcement and Recovery Act* (2009). SEC and other Banking Regulators routinely appear before the US Congress and provide testimonies, which claim that integrity of Banking Regulation in the United State has been "shored up".

The US Government was also pressured by the international community, which was seriously harmed by the ongoing financial crisis, to take corrective measures, consistent with the US Government's duties and obligations in international treaties and accords.

In contrast, the litigation of *SEC v BAC* in the US Court, Southern District of New York, reviewed in the current report, was as merely simulated litigation, and fraud on the People of the United States and the international community by the US District Court and the US government. This study also documents the tight link between integrity of the US courts, or lack thereof, and the failure of US banking regulation.

F. PACER and CM/ECF, Information Systems of the US Courts

The case, reviewed in the current study, documents the critical effect of PACER and CM/ECF, the electronic records systems of the US courts, on to integrity, or lack thereof, of US court records and the US justice system in general.

Conditions that today prevail in the US courts stand in stark contrast with the laws enacted by US Congress, Presidential Directives, Regulations, and applications, which have been presumably implemented for

 ISBN: 978-1-61208-242-4

authentication and validation of electronic government records in the United States. [26, 27]

Deficiencies in the design and operation of PACER and CM/ECF were previously noted in reports, which were published by others and by this author in legal, criminology, and computer science journals. [1, 8-10] The report in ProPublica, copied in the *National Law Journal* documented the falsification of records in a landmark case of international significance – the Habeas Corpus petition of a Guantanamo detainee in the US District Court, Washington DC. Such conduct should have been prevented, or made obvious by a valid electronic record system of the courts.

Reports regarding the deficiencies in PACER and CM/ECF have also been repeatedly forwarded to the appropriate US government agencies. [10] Regardless, there is no evidence of intent to initiate corrective actions.

G. *The NEFs as Certificates of Authentication – a Case of "Robo-signers"*

Of particular concern is the design and operation of the NEFs, the certificates of service in CM/ECF.

Beyond the design of the electronic form itself, as described above, the universal exclusion of the NEFs from public access has no plausible explanation that is consistent with integrity of the courts.

The opening statement on the NEF says:

> "This is an automatic e-mail message generated by the CM/ECF system."

Therefore, the overall design of the NEF, as a Certificate of Service, is comparable to systems, which have been reported in recent years in the banking system, as one of the causes of the financial crisis, and have been dubbed "Robo-signers".

The NEF is an electronic record that purports to provide certification of critical court records. In fact, it provides no no certification at all and implies accountability by no individual.

H. *The Electronic Documents Stamps – invalid either as a digital signatures or as a checksum strings*

As described above (Figures 1,2; see Online Appendix 1), the Electronic Document Stamp in the NEFs replaced the hand-signature of the Clerk of the Court in the Certificates of Service. However, the Stamp is a checksum string, and as such, cannot be deemed an electronic signature of any authorized, named individual. Moreover, the public is denied access to the NEFs.

The CM/ECF User Guide of the US District Court, Northern District of Illinois is unique among such user guides, which have been surveyed. It provides detailed description of a Document Verification Utility, which permits the CM/ECF user to verify a given PACER record against the Stamp in the respective NEF. (Figure 5; see Online Appendix 1) [28] However, the public is also denied access to the Document Verification Utility. Had the public been permitted access to the NEFs and the Document Verification Utility in the US District Court, Washington, DC, it would have provided definitive evidence for the adulteration of the Order in the Guantanamo Bay Habeas Corpus case, even absent to the two materially different order records, recently reported by the National Law Journal. [1]

There is no plausible explanation for such design of PACER and CM/ECF and conduct of the US courts in this regard that is consistent with integrity of the US courts.

I. *Data Mining of Judicial Records*

This study highlights the significance of data mining of judicial records and other records of the justice system (e.g., prisoners´ registration systems [8]). The integrity of these systems is critical for the safeguard of Human Rights and Civil Society, and their corruption bears a profound impact, which should be considered an unannounced regime change. This study also demonstrates that although the courts today deny access to key records, in violation of the law, through data mining, the fraud in such systems can still be elucidated.

Obviously, key data to be mined are elements related to graphic signatures, electronic signatures, authentication records, names and authorities of individuals involved in issuing the various records.

Beyond the documentation of the invalidity of specific records, data mining provides evidence of the invalidity and fraud in the design and operation of such systems as a whole.

J. *The fraud inherent in the electronic recod systems of the US courts is not unique.*

Sustain, the electronic record system of the Superior Court of California, County of Los Angeles, is believed to be one of the earliest electronic record system of any court (implemented around 1985). According to the web page of its maker, the Sustain Corporation, the system is by now implemented in the courts of eleven states and three nations. Sustain shows remarkably similar fraudulent features to those described here in PACER and CM/ECF.

An accompanying paper describes the electronic record systems of the courts of the State of Israel, which were implemented a decade later than PACER and CM/ECF. The Israeli systems also show remarkable similarity to the fraud inherent in PACER and CM/ECF.

Preliminary inspection suggests that similar faults also exist in the electronic record systems, which have been recently implemented in other "Western Democracies".

K. *Desired Corrective Actions*

Deficiencies in PACER and CM/ECF, which were outlined in this study, are not inherent to electronic records systems. On the contrary, such systems could have enhanced integrity and transparency of the courts and the judicial process.

 ISBN: 978-1-61208-242-4

1) *Computing Experts in General, and Data Mining Experts in Particular, Should Assume a Unique Civic Duty in the Ongoing Monitoring of the Integrity of the Electronic Record Systems of the Courts.*

The common law right to inspect and to copy judicial records was reaffirmed by the US Supreme Court in *Nixon v Warner Communications, Inc* (1978) as inherent to the First Amendment. In doing so, the US Supreme Court said that the right was necessary for the public "to keep a watchful eye on government". Today, the public must keep a watchful eye particularly on the courts' electronic record systems. Computing experts in general and data mining experts in particular are uniquely equipped to exert such "watchful eye" on the courts. No other measures could substitute for public scrutiny of the courts in safeguarding Human Rights and Civil Society in the Digital Era.

2) *Procedures Inherent to the Design and Operation of PACER and CM/ECF Should be Established by Law and Validated by Computing Experts under Public and Legal Accountability.*

Implementation of PACER and CM/ECF should be recognized as an act of establishing novel court procedures. Therefore, their implementation should have been established by law. Validation (functional logic verification) of such systems should have been undertaken prior to their installation, in a manner that is both legally and publicly accountable, e.g., through agencies under control of the legislative branch. As part of validation, adequate security should be ascertained, to ensure that only named, authorized court personnel is permitted to conduct docket transactions. The identity and authority, as well as digital signatures of such persons should be publicly and unambiguously discernable. Thereby, accountability of the clerks of the courts for integrity of electronic court records should be restored.

3) *Valid Authentication Procedures and Authentication Records Should be Implemented*

As detailed above, the NEFs cannot possibly be deemed valid electronic Certificate of Service, and the Electronic Document Stamp was implemented in CM/ECF in a manner that prevents the public from ascertaining the integrity, or lack thereof, of US court records. A publicly recognized and publicly accessible form of valid digital signatures should be implemented instead.

4) *Public Access to Judicial Records, to Inspect and to Copy, Should be Restored.*

Given that today PACER is effectively the exclusive way, provided by the US courts, for public access to court records, all court records, including the electronic Certificates of Service, should be publicly accessible online, pursuant to the First Amendment, Due Process, and Public Trial rights. Such access should be denied only when explicitly stipulated by law or through fully documented sealing orders.

5) *US judges and clerks should be required to post financial disclosures on an annual basis.*

Given the serious concern of bribing of judges and clerks by corporate and other financial interests, judges and clerks should be required to post their financial disclosures on an annual basis. Similar measures were instituted in California for various elected public officials, and also for police officers of the undercover narcotics unit of the Los Angeles Police Department, given substantial evidence of widespread public corruption.

V. CONCLUSIONS

Litigation of *SEC v BAC* (1:09-cv-06829) in the US District Court, Southern District of New York, was hailed by media as the hallmark of banking regulation in the wake of the global financial crisis. The current study finds critical records of the litigation missing, others - outright invalid, simulated records, and others - vague and ambiguous to the degree that the litigation as a whole cannot be reasonably deemed valid litigation. Instead, the evidence shows that the case was conducted as simulated litigation in the US District Court, Southern District of New York, from it onset. To a large degree, such conduct was enabled through the design and operation of PACER and CM/ECF, the electronic record systems of the US courts. Key defects in PACER and CM/ECF, which were demonstrated in the current report, pertain to digital signatures, authentication records, authorities, security, public access, and functional logic verification of the systems as a whole.

PACER and CM/ECF are invalid electronic records systems of the US courts.

This study is also a call for action by computing experts in general and data mining experts in particular for the safeguard of Human Rights and integrity of government in the Digital Era.

VI. ONLINE APPENDICES

[1] This complete paper, including Figures 2-5, is accessible at: http://www.scribd.com/doc/104880125/

[2] The complete PACER docket, additional relevant records in litigation of *SEC v BAC* (1:09-cv-06829), and detailed analysis of specific records are accessible at: http://www.scribd.com/doc/44663232/

VII. ACKNOWLEDGMENT

The author thanks Israeli computing/cryptology and legal experts for their assistance.

VIII. REFERENCES

[1] D. Linzer, "In GITMO Opinion, Two Versions of Reality", *The National Law Journal,* October 11, 2010.

[2] J. Zernik, 'Data Mining of Online Judicial Records of the Networked US Federal Courts', *1 International Journal on Social Media: Monitoring, Measurement, Mining* 1:69-83, 2010.

[3] A. Field, "Foreclosure fraud: The homeowner nightmares continue", CNN, April 7, 2011.

 ISBN: 978-1-61208-242-4

[4] Daily Koss, "60 Minutes Exposes Massive Foreclosure Fraud" April 11, 2011.

[5] G. Morgenson, "From East and West, Foreclosure Horror Stories", New York Times, January 7, 2012.

[6] G. Morgenson and L. Story, "In Financial Crisis, No Prosecutions of Top Figures", New York Times, April 14, 2011.

[7] S. Paltrow, "Special Report: The watchdogs that didn't bark," Reuters, December 11, 2011.

[8] The following paper details retaliation against the 70-year old, former US prosecutor Richard Fine, who exposed, publicized, and rebuked the large-scale bribing of the judges of the Superior Court of California, County of Los Angeles - the largest county court in the United States. His actions led to the signing of "retroactive immunities" (simulated pardons) to all judges of the California courts by the California Governor – in fact an admission of widespread criminality. Two weeks later, Richard Fine was arrested. He was falsely imprisoned for 18 months in solitary confinement (considered by the United Nations torture). His electronic prisoner's booking record listed him as arrested and booked on location and by authority of the "Municipal Court of San Pedro". Such court did not and does not exist. Richard Fine's Habeas Corpus petitions were subjected to simulated review in the US District Court, Central Disrict of California, the US Court of Appeals, 9th Circuit, and the US Supreme Court.

J. Zernik, "Habeas Corpus in the United States - the case of Richard Isaac Fine – a Review" (2010); retrieved: September 4, 2012; accessible at:

http://www.scribd.com/doc/24729084/

[9] The following papers detail simulated litigation in *Citizens United v Federal Election Commission* - a landmark case in the US District Court, Washington DC, and in the Supreme Court of the United States. The case is considered by some legal experts as an open invitation by the US Supreme Court for the corruption of the US governemnt. The Supreme Court docket in this case lists "judgment issued", but no judgment record is to be found in the online electronic records of the Supreme Court and no judgment record was found by the Federal Election Commission in response to *Freedom of Information Act* request.

J. Zernik, "*Citizens United v Federal Election Commission* (1-07-cv-2240) in the US District Court, DC – invalid court records in a case of Simulated Litigation" (2011); retrieved: September 4, 2012; accessible at:

http://www.scribd.com/doc/56080106/

J. Zernik, "*Citizens United v Federal Election Commission* in the US Supreme Court - so far only a simulated Judgment record has been discovered…" (2011); retrieved: September 4, 2012; accessible at:

http://www.scribd.com/doc/55613401/

[10] The following papers includes a list of US judges and respective cases of simulated litigation in the US district courts, the US courts of appeals, and the US Supreme Court, with links to detailed review of each:

W. Windsor and J. Zernik, "Request filed by Windsor and Zernik with US Attorney General Eric Holder for Review of Integrity of Public Access and Case Management Systems of the US Courts" (2011); retrieved: September 4, 2012; accessible at:

http://www.scribd.com/doc/59480718/

J. Zernik, "Evidence of widespread corruption of the US courts and proposed corrective measures submitted to the US House of Representatives"; retrieved: September 4, 2012; accessible at:

http://www.scribd.com/doc/85481555/

[11] J. Bohm, "Memorandum Opinion", in *Case of Borrower Parsley* (05-90374), US Bankruptcy Court, Southern District of Texas, Dkt #248, March 5, 2008; retrieved: September 4, 2012; accessible at:

http://www.scribd.com/doc/25001966/

[12] P. Carrizosa, "RE: Engagement of Attorney McCormick in *Fine v Sheriff* (2:09-cv-01914) in the US District Court, Central District of California", March 19, 2010 Letter on behalf of the California Judicial Council"; retrieved: September 4, 2012; accessible at: http://www.scribd.com/doc/28645522/

[13] The terms "Simulated litigation", "simulated decisions", "simulated service", etc are used in this paper as defined in the Texas Penal Code:

> Texas Penal Code §32.48. SIMULATING LEGAL PROCESS.
> (a) A person commits an offense if the person recklessly causes to be delivered to another any document that simulates a summons, complaint, judgment, or other court process with the intent to:
> (1) induce payment of a claim from another person; or
> (2) cause another to:
> (A) submit to the putative authority of the document; or
> (B) take any action or refrain from taking any action in response to the document, in compliance with the document, or on the basis of the document.
> (b) Proof that the document was mailed to any person with the intent that it be forwarded to the intended recipient is a sufficient showing that the document was delivered.

The same conduct is defined in the penal codes of some other states as "sham litigation". In the US Code the same conduct is prohibited under the *Racketeer Influenced and Corrupt Organization Act* (RICO).

[14] US District Court, Southern District of New York, "Local Rules and ECF Rules and Instructions" (2010).

[15] W. Anderson, "Anderson on Civil E-Filing in the Central District of California: The Unofficial Civil E-filing User Manual", 2008.

[16] Securities and Exchange Commission, "September 9, 2010 FOIA Response No 10-03964-FOIA, RE: *SEC v Bank of America Corporation* (1:09-cv-06829)"; retrieved September 4, 2012; accessible at:

http://www.scribd.com/doc/46608559/

[17] L. Story, "Judge Accepts S.E.C.'s Deal With Bank of America", *New York Times*, February 22, 2010.

[18] Judicial Conference of the United States, Committee on Court Administration and Case Management, "Civil Litigation Management Manual", 2nd Edition (2010).

[19] A. Cuomo, "Re: Bank of America - Merrill Lynch Merger Investigation," April 23, 2009, linked to: D. Cho and T. Tse "U.S. Forced Bank Board To Carry Out Merrill Deal," Washington Post, April 24, 2009

[20] A. Jones, "Rakoff Hands It to BofA, the SEC,", *Law Blog*, *The Wall Street Journal,* August 11, 2009

[21] The term "show trials" originally referred to certain trials in the Soviet Union. Show trials are described in Wikipedia as follows:

> The term **show trial** is a pejorative description of a type of highly public trial in which there is a strong connotation that the judicial authorities have already determined the guilt of the defendant. The actual trial has as its only goal to present the accusation and the verdict to the public as an impressive example and as a warning.

 ISBN: 978-1-61208-242-4

Show trials tend to be retributive rather than correctional justice. The term was first recorded in the 1930s.

The term "inverse show trial" here refers to the practice in the US today - conducting simulated litigation in cases, where the goal is to publicly exonerate certain individuals after their criminality has been publicly exposed, e.g., government officers and bankers.

[22] J. Zernik, "Motion to Intervene and Concomitantly Filed Papers," in *Log Cabin Republicans v USA et al* (10-56634) in the US Court of Appeals, 9th Circuit Dkt #39-47, January 7, 2011; retrieved September 4, 2012; accessible at:
http://www.scribd.com/doc/46516034/

[23] J. Zernik, "Amended Request for Correction of US Supreme Court Records" in *Fine v Sheriff* (09-A827); retrieved September 4, 2012; accessible at:
http://www.scribd.com/doc/30162109/

[24] In *Bulloch v. United States*, 763 F.2d 1115, 1121 (10th Cir. 1985), the court states "Fraud upon the court is fraud which is directed to the judicial machinery itself and is not fraud between the parties or fraudulent documents, false statements or perjury. ... It is where the court or a member is corrupted or influenced or influence is attempted or where the judge has not performed his judicial function --- thus where the impartial functions of the court have been directly corrupted."

[25] J. Zernik, "Through Implementation of Case Management Systems, the California and the US Courts Rendered the Judicial Process Vague and Ambiguous"; retrieved September 4, 2012; accessible at:
http://www.scribd.com/doc/46519282/

[26] *E-Sign Act* (2000), USA

[27] *E-Government Act* (2002), USA

[28] US District Court, Northern District of Illinois, "CM/ECF Ver4.0 User Guide", undated.

[29] United Nations Drug Control and Crime Prevention Center, "Report of the First Vienna Convention - Strengthening Judicial Integrity, CICP-6" (2000).

[30] United Nations Drug Control and Crime Prevention Center, "Strengthening Judicial Integrity Against Corruption CICP-10" (2001)

 ISBN: 978-1-61208-242-4

The Open Data Interface (ODI) Framework for Public Utilization of Big Data

Hwa-Jong Kim
Dept. Computer Engineering
Kangwon National University
ChunCheon, Korea, 200-701
hjkim3@gmail.com

Seung-Teak Lee
IT Convergence Service Dept.
National Information society Agency
(NIA), Seoul, Korea
leest@nia.or.kr

Yi-Chul Kang
IT Convergence Service Dept.
National Information society Agency
(NIA), Seoul, Korea
kangyc@nia.or.kr

***Abstract*—In the paper, the Open Data Interface (ODI) framework for public utilization of Big Data was suggested. Comparing to conventional Web based open APIs which provide restricted access to the internal database of large data companies, the ODI provides multi-level access of data in public domain, ranging from copying plain files to extracting executive summary. Through the ODI, users can share raw data, intermediate mined data, or graphical reports, and can contribute to public utilization of Big Data. In the National Information society Agency (NIA) of Korea, the ODI scheme is considered as a public data infrastructure to support big and small companies for future data mash up business. Many companies in Korea in the field of telecommunications and web services are interested in developing a collaborative public data infrastructure with the guide of government. In the paper, we proposed an initial test infrastructure for the purpose.**

Keywords-big data；open API；open data interface； public data

I. INTRODUCTION

These days, Big Data is attracting much attention because of its potential benefits to find valuable information from plain data. Big Data services include data gathering, data analysis, data mining, recommendation, prediction, and reporting by using various data sources such as, sales data, social network service messages, location information, and any related documents.

However, together with the prospecting advantages of Big Data, some problems are also expected in the Big Data world:

- Monopolization of data by large data companies
- Digital divide in data accessibility
- Big data traffic due to redundant copies

As the Big Data service is growing, a few large *data companies* such as Google, Facebook, Twitter, Amazon, Yahoo, Naver, or Daum (big web portals in Korea) will have a good chance of gathering valuable data every day. The large data companies can make use of the big data for marketing and service improvement, which again gathers more valuable data from the users. This will give more severe digital divide in data access capabilities for individuals and small companies.

The Big Data service will eventually incur big traffics. The volume of data in the world will grow, and as far as many Big Data services are introduced, a transformed data set will also be generated by many companies, institutes, and individuals. This will also produce redundant data set which is almost same to the original data except that only a very small part is changed. The main bottleneck of Big Data may come from the telecommunication channel rather than memory or computing resources. The bandwidth of channel is always limited, and costs high.

In order to handle above problems, we suggest the Open Data Interface (ODI) framework for public utilization of Big Data. The ODI is an extended framework of the conventional Web based open application program interface (API), but providing more flexible and unconstrained access interface. With the ODI, users can access raw data file, intermediate mined data, or graphical reports. With the conventional Web based APIs, users can get data from sites, but cannot put his or her processed (or mined) data to the sites. In other words, API users cannot contribute to build a public data infrastructure with a more rich set of raw data or mined data. With the ODI, users can put related data, processed data or even mining algorithms to the ODI infrastructure.

In the paper, we briefly review related work, and describe the concepts of the ODI framework and its operations.

II. RELATED WORK

Many companies (or institutes) provide Web based open APIs for users to access database in the company. There are more than 6000 sets of APIs, including Twitter, YouTube, Facebook, Google Maps, Flickr, LinkedIn, etc. [1]. But the open APIs have the following limitations (see Figure 1):

- Open APIs provide limited access to data in size and types
- Open API requires professional programming skill
- the purpose of the open API is for the company, and activity history is accumulated in the company

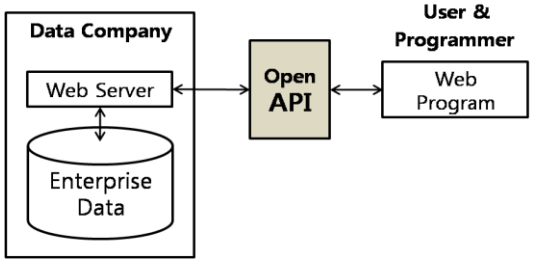


Figure 1. With conventional open API, the users should develop high-tech web programs to get data from the database of the company.

 ISBN: 978-1-61208-242-4

Even though the data company provides open API, it is usually limited in the scope and volume of the access data. Furthermore in order to use the open API, we need skilled programming. It is noted that the final purpose of the open API is for the company. It is not for public good.

Recently, it is suggested that Big Data should support improving public services in health, safety, and creating new business [2]. For this purpose, a public data space will be needed for easy sharing of data.

A related work is the Linked Data which was proposed to connect many types of data over the Web by using the Uniform Resource Identifiers (URIs), HTTP for identification, and Resource Description Framework (RDF) for contents description [3]. The Linking Open Data (LOD) project [4] uses the Web like a single global database to integrate data from heterogeneous sources. The LOD helps users navigate between related data sets through the semantic web. Some challenges of Big Data were investigated including data quality and lack of good use cases [5].

The Interaction design during the process of acquiring, analyzing, and using the Big Data is also becoming critical issue for success of Big Data [6]. Machine learning algorithms can be applied to large data sets over hadoop platform [7], and a cloud based prediction service is provided by Google [8].

III. DEFINITION OF THE ODI FRAMEWORK

A. Background of the ODI

The LOD was proposed to link related data over the Web, and therefore can be used for public Big Data service because Web is open to anyone. However, the LOD is mainly focused on connecting related data sets and finding them efficiently. But we need a more flexible framework which can integrate, besides the data itself, the machine learning algorithms used, and specific domain knowledge obtained from the case.

In the proposed ODI framework, we expect distributed contribution of users in processing (e.g., data mining) and utilizing the raw and intermediate mined data. The ODI provides a multi-level access and processing of information based on closely related data sets by many contributors.

B. Public Data Space (PDS)

The ODI framework model is shown in Figure 2, where the ODI is used to access the Public Data Space (PDS) by many contributors from government, enterprise and individuals.

The PDS is composed of Data Core and Contributed Local Data. Data Core is a marketplace where every data can be searched and accessed and processed. It includes some core public data. Contributed Local Data is data residing in the users' server, such as government (public data), enterprise (open data for the public), and individuals (privately processed public data). The PDS is a platform of sharing data, and can be regarded as a collaborative data warehouse.

Each user may have its own private data that is not shared with others.

C. Data Core

Data Core is composed of physical storage of data, virtual collection of data which is physically located in the user servers, and data analytics functions. Data Core also provides programmable platform which include gathering, analysis, and reporting functions. In other words, Data Core is composed of data and functions contributed by any users.

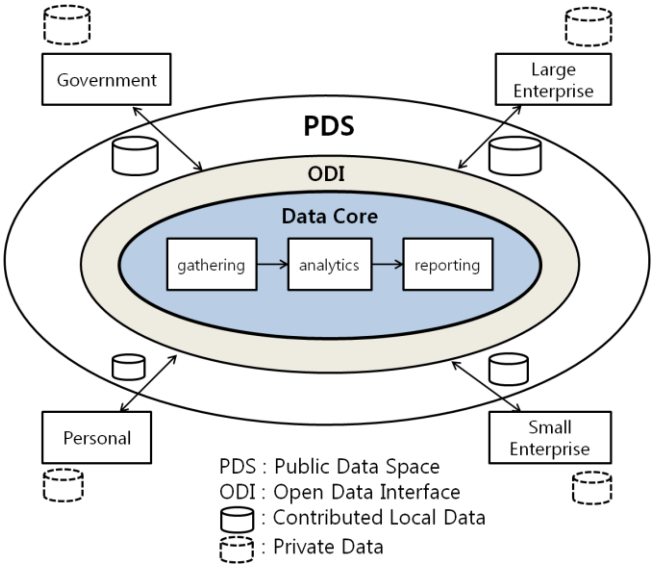


Figure 2. The ODI framework provides gathering, analytics, and reporting operations to the PDS

D. Open Data Interface (ODI)

The ODI uses the Data Core (its data and function) in order to provide various types of interface for gathering, analytics, and reporting of data. The goal of ODI is to provide easy but standardized interface in accessing data and sharing domain knowledge. We will explain the operations of the ODI framework.

IV. OPERATIONS OF THE ODI

A. Multi-level accessing

The ODI provides multi-level accessing to the PDS. The ODI provides various types of APIs to access raw data, processed data, abstract data, and they also do some operations of gathering, analyzing, and reporting. In other words, the ODI provides plain file copying, running machine learning algorithms, executive summary, or a graphic processing (see Figure 3).

In the ODI model, expert programmers are involved in developing the APIs in the Data Core and interface libraries in the ODI shell. Plain users may just input commands to the ODI to get some data or result.

 ISBN: 978-1-61208-242-4

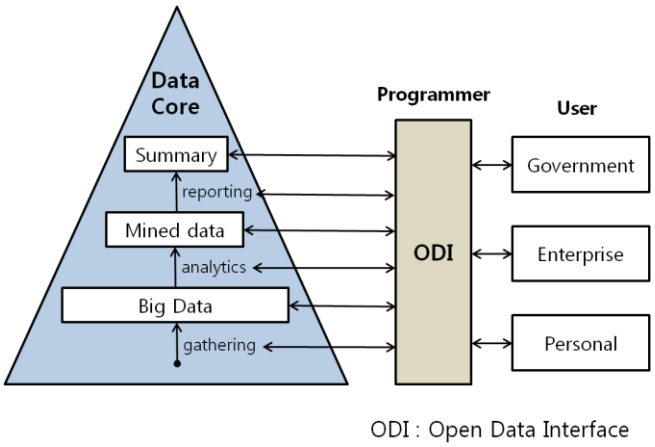


Figure 3. With ODI, the users need not to run web programs, but can get multi-level access to data and functions on the PDS.

B. *Functional Requirement of the ODI*

The ODI provides simple and standardized access to the Data Core. The Data Core is composed of data and functions operating on the data. The ODI should categorize the levels of accessing, i.e., levels of abstraction and processing levels. There are two types of levels:

- Data level (e.g., raw, processed, summarized data)
- Functional level (e.g., gathering, analysis, reporting)

The ODI interprets the response from Data Core to the users, and separates the role of user from being a professional programmer

V. USE OF THE ODI

Many companies and governments are gathering huge and valuable data in their domain every day, but do not fully utilize the data. It is because the data is isolated. For example, telecommunication companies may need banking information or Internet search history of their users for more intelligent services. In the future, new disruptive services will come from data mash up among various types of companies and government's public data. The ODI will help the extension of data mash up.

For data mash up cooperation among companies, we need an open infrastructure where each company can give (put) and get beneficial data in a standard and safe way. They want to sell processed (or screened) data and buy their missing data in an open market. The government should help the operations of the market through standardization of data format and access rules. We also need regulations in privacy-preserving data mining.

With the ODI, users can share their domain knowledge in the form of mined data or a new algorithm. For example, we can get top 20 news from a news portal. If an expert classified the top 20 news in an interesting way, he or she can share the idea by putting back the processed data to the news portal with the newly developed APIs. With this processed (or mined) data, other users may save time or memory by avoiding the same analysis.

We also expect the ODI may alleviate the drastic increase data traffic due to big data applications. The ODI will minimize the redundant copy of similar data by redundant (or similarity) checking. Traditionally, the usefulness of data mainly depends on the correctness of data. But in Big Data, the usefulness of data will mainly depend on timely access of data because the data preparation takes long time. For example, if two unstructured data set (e.g. blog data or news data) differs only by 1%, they may be regarded as a same data, and does not need to transfer them again for perfect coincidence.

VI. CONCLUSION AND FUTURE WORK

In the paper, we introduced the ODI framework which can be used to for public Big Data application by providing nationwide PDS infrastructure. The PDS and ODI framework need to be installed and operated by the government for public good, and minimizing digital divide in the coming Big Data era. The ODI framework is for easy access of data, algorithms and sharing success cases. It provides a kind of public data warehousing with related machine learning algorithms proven to be useful for some applications.

In the ODI model, the quality of data is not measured by the accuracy but by the usefulness of the data, which is evaluated by the users. We also hope that the ODI is used by individuals or small companies who want to create a new business in the Big Data world.

ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0301-12-1004)

REFERENCES

[1] http://www.programmableweb.com/ [retrieved: June, 2012]

[2] Alex Howard, Data for the Public Good, O'Reilly Media, 2012.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far" International Journal on Semantic Web & Information Systems, Vol. 5, Iss. 3, pp. 1-22, 2009.

[4] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[5] Christian Bizer, Peter Boncz, Michael L. Brodie, Orri Erling, "The meaningful use of big data: four perspectives -- four challenges", ACM SIGMOD Record, Vol. 40, Iss. 4, pp. 56-60, 2012.

[6] Danyel Fisher, Rob DeLine, Mary Czerwinski, Steven Drucker, "Interactions with big data analytics", Interactions, Vol. 19, Iss. 3, pp. 50-59, 2012.

[7] Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman, Mahout in Action, Manning, 2011.

[8] https://developers.google.com/prediction/ [retrieved: June, 2012]

 ISBN: 978-1-61208-242-4

Evaluating Data Minability Through Compression – An Experimental Study

Dan Simovici
Univ. of Massachusetts Boston,
Boston, USA,
dsim at cs.umb.edu

Dan Pletea
Univ. of Massachusetts Boston,
Boston, USA,
dpletea at cs.umb.edu

Saaid Baraty
Univ. of Massachusetts Boston,
Boston, USA,
sbaraty at cs.umb.edu

***Abstract*—The effectiveness of compression algorithms is increasing as the data subjected to compression contains repetitive patterns. This basic idea is used to detect the existence of regularities in various types of data ranging from market basket data to undirected graphs. The results are quite independent of the particular algorithms used for compression and offer an indication of the potential of discovering patterns in data before the actual mining process takes place.**

***Keywords*-data mining; lossless compression; LZW; market basket data; patterns; Kronecker product.**

I. INTRODUCTION

Our goal is to show that compression can be used as a tool to evaluate the potential of a data set of producing interesting results in a data mining process. The basic idea that data that displays repetitive patterns or patterns that occur with a certain regularity will be compressed more efficiently compared to data that has no such characteristics. Thus, a pre-processing phase of the mining process should allow to decide whether a data set is worth mining, or compare the interestingness of applying mining algorithms to several data sets.

Since compression is generally inexpensive and compression methods are well-studied and understood, pre-mining using compression will help data mining analysts to focus their efforts on mining resources that can provide a highest payout without an exorbitant cost.

Compression has received lots of attention in the data mining literature. As observed by Mannila [7], data compression can be regarded as one of the fundamental approaches to data mining [7], since the goal of the data mining is to "compress data by finding some structure in it".

The role of compression developing parameter-free data mining algorithms in anomaly detection, classification and clustering was examined in [4]. The size $C(x)$ of a compressed file x is as an approximation of Kolmogorov complexity [2] and allows the definition of a pseudo-distance between two files x and y as

$$d(x,y) = \frac{C(xy)}{C(x)+C(y)}.$$

Further advances in this direction were developed in [8][5][6]. A Kolmogorov complexity-based dissimilarity was successfully used to texture matching problems in [1] which have a broad spectrum of applications in areas like bioinformatics, natural languages. and music.

We illustrate the use of lossless compression in pre-mining data by focusing on several distinct data mining processes: files with frequent patterns, frequent itemsets in market basket data, and exploring similarity of graphs.

The LZW (Lempel-Ziv-Welch) algorithm was introduced in 1984 by T. Welch in [9] and is among the most popular compression techniques. The algorithm does not need to check all the data before starting the compression and the performance is based on the number of the repetitions and the lengths of the strings and the ratio of 0s/1s or true/false at the bit level. There are several versions of the LZW algorithm. Popular programs (such as Winzip) use variations of the LZW compression. The Winzip/Zip type of algorithms also work at the bit level and not at a charater/byte level.

We explore three experimental settings that provide strong empirical evidence of the correlation between compression ratio and the existence of hidden patterns in data. In Section II, we compress binary strings that contain patterns; in Section III, we study the compressibility of adjacency matrix for graphs relative to the entropy of distribution of subgraphs. Finally, in Section IV, we examine the compressibility of files that contain market basket data sets.

II. PATTERNS IN STRINGS AND COMPRESSION

Let A^* be the set of strings on the alphabet A. The length of a string w is denoted by $|w|$. The null string on A is denoted by λ and we define A^+ as $A^+ = A^* - \{\lambda\}$.

If $w \in A^*$ can be written as $w = utv$, where $u, v \in A^*$ and $t \in A^+$, we say that the pair (t, m) is an occurrence of t in w, where m is the length of u.

The occurrences (x, m) and (y, p) are overlapping if $p < m + |x|$. If this is the case, there is a proper suffix of x that equals a proper prefix of y. If t is a word such that the sets of its proper prefixes and its proper suffixes are disjoint, there are no overlapping occurrences of x in any word. The number of occurrences of a string t in a string w is denoted by $n_t(w)$. Clearly, we have $\sum\{n_a(w) \mid a \in A\} = |w|$. The *prevalence* of t in w is the number $f_t(w) = \frac{n_t(w)\cdot|t|}{|w|}$ which gives the ratio of the characters contained in the occurrences of t relative to the total number of characters in the string.

 ISBN: 978-1-61208-242-4

The result of applying a compression algorithm C to a string $w \in A^*$ is denoted by $C(w)$ and the *compression ratio* is the number

$$\mathsf{CR}_C(w) = \frac{|C(w)|}{|w|}.$$

In this section, we shall use the binary alphabet $B = \{0, 1\}$ and the LZW algorithm or the compression algorithm of the package `java.util.zip`.

We generated random strings of bits (0s and 1s) and computed the compression ratio strings with a variety of symbol distributions. A string w that contains only 0s (or only 1s) achieves a very good compression ratio of $\mathsf{CR}_{jZIP}(w) = 0.012$ for 100K bits and $\mathsf{CR}_{jZIP} = 0.003$ for 500K bits, where $jZIP$ denotes the compression algorithm from the package `java.util.zip`. Figure 1 shows, as expected, that the worst compression ratio is achieved when 0s and 1s occur with equal frequencies.

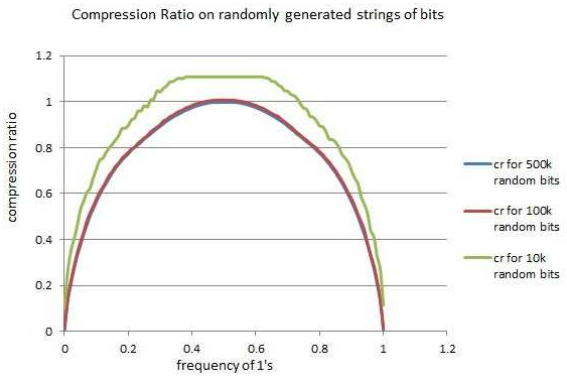


Figure 1. Baseline CR_{jZIP} Behavior

For strings of small length (less than 10^4 bits) the compression ratio may exceed 1 because of the overhead introduced by the algorithm. However, when the size of the random string exceeds 10^6 bits this phenomenon disappears and the compression ratio depends only on the prevalence of the bits and is relatively independent on the size of the file. Thus, in Figure 1, the curves that correspond to files of size 10^6 and $5 \cdot 10^6$ overlap. We refer to the compression ratio of a random string w with an $(n_0(w), n_1(w))$ distribution as the *baseline compression ratio*.

We created a series of binary strings $\varphi_{t,m}$ which have a minimum guaranteed number m of occurrences of patterns $t \in \{0,1\}^k$, where $0 \leq m \leq 100$. Specifically, we created 101 files $\varphi_{001,m}$ for the pattern 001, each containing 100K bits and we generated similar series for $t \in \{01, 0010, 00010\}$. The compression ratio is shown in Figure 2. The compression ratio starts at a value of 0.94 and after the prevalence of the pattern becomes more frequent than 20% the compression ratio drops dramatically. Results of the experiment are shown in Table I and in Figure 3.

Table I
PATTERN '001' PREVALENCE VERSUS THE CR_{jZIP}

Prevalence of '001' pattern	CR_{jZIP}	Baseline
0%	0.93	0.93
10%	0.97	0.93
20%	0.96	0.93
30%	0.92	0.93
40%	0.86	0.93
50%	0.80	0.93
60%	0.72	0.93
70%	0.62	0.93
80%	0.48	0.93
90%	0.31	0.93
95%	0.19	0.93
100%	0.01	0.93

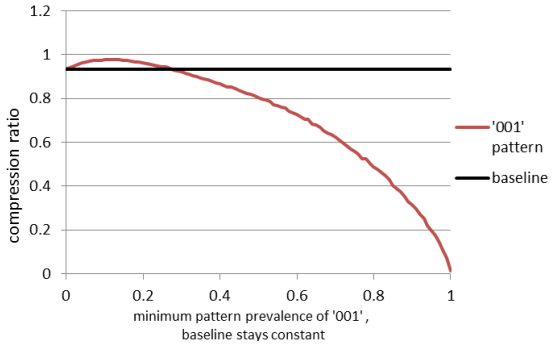


Figure 2. Variation of compression rate depends on the prevalence of the pattern '001'

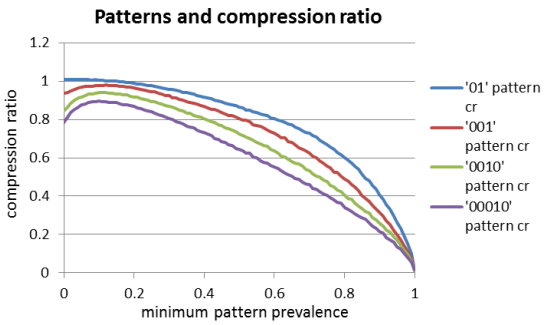


Figure 3. Dependency of Compression Ratio on Pattern Prevalence

We conclude that the presence of repeated patterns in strings leads to a high degree of compression (that is, to low compression ratios). Thus, a low compression ratio for a file indicates that the mining process may produce interesting results.

III. RANDOM INSERTION AND COMPRESSION

For a matrix $M \in \{0,1\}^{u \times v}$ denote by $n_i(M)$ the number of entries of M that equal i, where $i \in \{0, 1\}$. Clearly, we have $n_0(B) + n_1(B) = uv$. For a random variable V which

ranges over the set of matrices $\{0,1\}^{u\times v}$ let $\nu_i(V)$ be the random variable whose values equal the number of entries of V that equal i, where $i \in \{0,1\}$.

Let $A \in \{0,1\}^{p\times q}$ be a $0/1$ matrix and let

$$\mathcal{B} : \begin{pmatrix} B_1 & B_2 & \cdots & B_k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix},$$

be a matrix-valued random variable where $B_j \in \mathbb{R}^{r\times s}$, $p_j \geq 0$ for $1 \leq j \leq k$, and $\sum_{j=1}^{k} p_j = 1$.

Definition 3.1: The random variable $A \leftarrow \mathcal{B}$ obtained by the *insertion* of $\mathcal{B}$ into A is given by

$$A \otimes \mathcal{B} = \begin{pmatrix} a_{11}\mathcal{B} & \ldots & a_{1n}\mathcal{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathcal{B} & \ldots & a_{mn}\mathcal{B} \end{pmatrix} \in \mathbb{R}^{mr\times ns}$$

In other words, the entries of $A \leftarrow \mathcal{B}$ are obtained by substituting the block $a_{ij}B_\ell$ with the probability p_ℓ for a_{ij} in A. □

Note that this operation is a probabilistic generalization of Kronecker's product for if

$$\mathcal{B} : \begin{pmatrix} B_1 \\ 1 \end{pmatrix},$$

then $A \leftarrow B$ has as its unique value the Kronecker product $A \otimes B$.

The expected number of 1s in the insertion $A \leftarrow \mathcal{B}$ is

$$E[\nu_1(A \leftarrow \mathcal{B})] = n_1(A) \sum_{j=1}^{k} n_1(B_j)p_j$$

When $n_1(B_1) = \cdots = n_1(B_k) = n$, we have $E[\nu_1(A \leftarrow \mathcal{B})] = n_1(A)n$.

In the experiment that involves insertion, we used a matrix-valued random variable such that $n_1(B_1) = \cdots = n_1(B_k) = n$. Thus, the variability of the values of $A \leftarrow \mathcal{B}$ is caused by the variability of the contents of the matrices $B_1, \ldots, B_k$ which can be evaluated using the entropy of the distribution of $\mathcal{B}$,

$$\mathcal{H}(\mathcal{B}) = -\sum_{j=1}^{k} p_j \log_2 p_j.$$

We expect to obtain a strong positive correlation between the entropy of $\mathcal{B}$ and the degree of compression achieved on the file that represents the matrix $A \leftarrow \mathcal{B}$, and the experiments support this expectation.

In a first series of compressions, we worked with a matrix $A \in \{0,1\}^{106\times 106}$ and with a matrix-valued random variable

$$\mathcal{B} : \begin{pmatrix} B_1 & B_2 & B_3 \\ p_1 & p_2 & p_3 \end{pmatrix},$$

where $B_j \in \{0,1\}^{3\times 3}$, and $n_1(B_1) = n_1(B_2) = n_1(B_3) = 4$. Several probability distributions were considered, as shown in Table II. Values of $A \leftarrow \mathcal{B}$ had $106^2 * 3^2 = 101124$ entries.

In Table II, we had 39% 1s and the baseline compression rate for a binary file with this ratio of 1s is 0.9775. We also computed the correlation between the CR_{jZIP} and the Shannon entropy of the probability distribution and obtained the value 0.9825 for 3 matrices. In Table III, we did the same experiment but with 4 different matrices of 4×4. A strong correlation (0.992.) was observed between CR_{jZIP} and the Shannon entropy of the probability distribution.

Table II
MATRIX INSERTIONS, ENTROPY AND COMPRESSION RATIOS

Probability distribution	CR_{jZIP}	Shannon Entropy
$(0,1,0)$	0.33	0
$(1,0,0)$	0.33	0
$(0,0,1)$	0.33	0
$(0.2,0.2,0.6)$	0.77	1.37
$(0.6,0.2,0.2)$	0.74	1.37
$(0.33,0.33,0.34)$	0.79	1.58
$(0,0.3,0.7)$	0.7	0.88
$(0.9,0.1,0)$	0.51	0.46
$(0.8,0,0.2)$	0.61	0.72
$(0.49,0.25,0.26)$	0.77	1.5
$(0.15,0.35,0.5)$	0.78	1.44

Table III
KRONECKER PRODUCT AND PROBABILITY DISTRIBUTION FOR 4 MATRICES

Probability distribution	CR_{jZIP}	Shannon Entropy
$(0,1,0,0)$	0.23	0
$(0,1,0,0)$	0.23	0
$(0.2,0.2,0.2,0.4)$	0.69	01.92
$(0.25,0.25,0.25,0.25)$	0.69	2
$(0.4,0,0.2,0.4)$	0.53	1.52
$(0.3,0.1,0.2,0.4)$	0.65	1.84
$(0.45,0.12,0.22,0.21)$	0.61	1.83

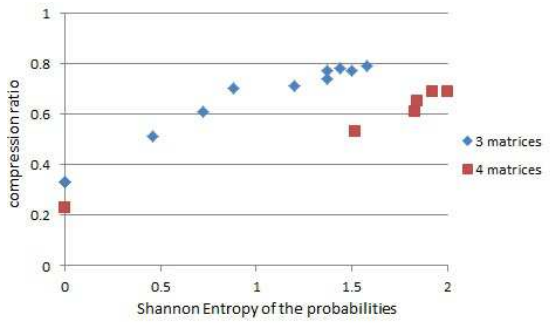


Figure 4. Evolution CR_{jZIP} and Shannon Entropy of Probability Distribution.

In Figure 4, we have the evolution of CR_{jZIP} on the y axis and on the x axis the Shannon Entropy of the probability distribution for both experiments. We can see clearly the linear correlation between the two.

 ISBN: 978-1-61208-242-4

This experiment proves us again that in case of repetitions/patterns the CR_{jZIP} is better than in the case of randomly generated files.

Next, we examine the compressibility of binary square matrices and its relationship with the distribution of principal submatrices. A binary square matrix is compressed by first vectorizing the matrix and then compressing the binary sequence. The issue is relevant in graph theory, where the principal submatrices of the adjacency matrix of a graph correspond to the adjacency matrices of the subgraphs of that graph. The patterns in a graph are captured in the form of frequent isomorphic subgraphs.

There is a strong correlation between the compression ratio of the adjacency matrix of a graph and the frequencies of the occurrences of isomorphic subgraphs of it. Specifically, the lower the compression ratio is, the higher are the frequencies of isomorphic subgraphs and hence the worthier is the graph for being mined.

Let $\mathcal{G}_n$ be an undirected graph having $\{v_1, \ldots, v_n\}$ as its set of nodes. The adjacency matrix of $\mathcal{G}_n$, $\mathbf{A}_{\mathcal{G}_n} \in \{0,1\}^{n\times n}$ is defined as

$$(\mathbf{A}_{\mathcal{G}_n})_{ij} = \begin{cases} 1 & \text{if there is an edge between } v_i \text{ and } v_j \text{ in } \mathcal{G}_n \\ 0 & \text{otherwise.} \end{cases}$$

We denote with $\mathsf{CR}_C(\mathbf{A}_{\mathcal{G}_n})$ the compression ratio of the adjacency matrix of graph $\mathcal{G}_n$ obtained by applying the compression algorithm C. Define the *principal subcomponent* of matrix $\mathbf{A}_{\mathcal{G}_n}$ with respect to the set of indices $S = \{s_1, \ldots, s_k\} \subseteq \{1, 2, \ldots, n\}$ to be the $k \times k$ matrix $\mathbf{A}_{\mathcal{G}_n}(S)$ such that

$$\mathbf{A}_{\mathcal{G}_n}(S)_{ij} = \begin{cases} 1 & \text{if there is an edge between } v_{s_i} \text{ and } v_{s_j} \\ & \text{in } \mathcal{G}_n \\ 0 & \text{otherwise.} \end{cases}$$

The matrix $\mathbf{A}_{\mathcal{G}_n}(S)$ is the adjacency matrix of the subgraph of $\mathcal{G}_n$ which consists of the nodes with indices in S along with those edges that connect these nodes. We denote by $\mathcal{P}_n(k)$ the collection of all subsets of $\{1, 2, \ldots, n\}$ of size k where $2 \leq k \leq n$. We have $|\mathcal{P}_n(k)| = \binom{n}{k}$.

Let $(\mathbf{M}_1^k, \ldots, \mathbf{M}_{\ell_k}^k)$ be an enumeration of possible adjacency matrices of graphs with k nodes where $\ell_k = 2^{\frac{k(k-1)}{2}}$. We define the finite probability distribution

$$P(\mathcal{G}_n, k) = \left(\frac{\mathsf{n}_1^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|}, \ldots, \frac{\mathsf{n}_{\ell_k}^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|} \right),$$

where $\mathsf{n}_i^k(\mathcal{G}_n)$ for $1 \leq i \leq \ell_k$ is the number of subgraphs of $\mathcal{G}_n$ with adjacency matrix $\mathbf{M}_i^k$. The Shannon entropy of this probability distribution is:

$$\mathcal{H}_P(\mathcal{G}_n, k) = -\sum_{i=1}^{\ell_k} \frac{\mathsf{n}_i^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|} \log_2 \frac{\mathsf{n}_i^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|}.$$

If $\mathcal{H}_P(\mathcal{G}_n, k)$ is low, there are to be fewer and larger sets of isomorphic subgraphs of $\mathcal{G}_n$ of size k. In other words, small values of $\mathcal{H}_P(\mathcal{G}_n, k)$ for various values of k suggest that the graph $\mathcal{G}_n$ contains repeated patterns and is susceptible to produce interesting results. Note that although two isomorphic subgraphs do not necessarily have the same adjacency matrix, the number $\mathcal{H}_P(\mathcal{G}_n, k)$ is a good indicator of the frequency of isomorphic subgraphs and hence subgraph patterns.

We evaluated the correlation between $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ and $\mathcal{H}_P(\mathcal{G}_n, k)$ for different values of k.

As expected, the compression ratio of the adjacency matrix and the distribution entropy of graphs are roughly the same for isomorphic graphs, so both numbers are characteristic for an isomorphism type. If ϕ is a permutation of the vertices of $\mathcal{G}_n$, the adjacency matrix of the graph $\mathcal{G}_n^\phi$ obtained by applying the permutation is defined by $\mathbf{A}_{\mathcal{G}_n^\phi}$ is given by

$$\mathbf{A}_{\mathcal{G}_n^\phi} = P_\phi \mathbf{A}_{\mathcal{G}_n} P_\phi^{-1}.$$

We compute this adjacency matrix of $\mathbf{A}_{\mathcal{G}_n^\phi}$, the entropy $\mathcal{H}_P(\mathcal{G}_n^\phi, k)$ the compression ratio $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n^\phi})$ for several values of k and permutations.

We randomly generated graphs with $n = 60$ nodes and various number of edges ranging from 5 to 1765. For each generated graph, we randomly produced twenty permutations of its set of nodes and computed $\mathcal{H}_P(\mathcal{G}_n^\phi, k)$ and $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n^\phi})$.

Finally, for each graph we calculated the ratio of standard deviation over average for the computed compression ratios, followed by the same computation for distribution entropies.

The results of this experiment are shown in Figures 5 and 6 against the number of edges. As it can be seen, the deviation over mean of the compression ratios for $n = 60$ does not exceed the number 0.05. Also, the deviation over average of the distribution entropies for various values of k do not exceed 0.006. In particular, the deviation of the distribution entropy for the graphs of 100 to 1500 edges falls below 0.001, which allows us to conclude that the deviations of both compression ratio and distribution entropy with respect to isomorphisms are negligible.

For each $k \in \{3, 4, 5\}$, we generated randomly 560 graphs having 60 vertices and sets of edges whose size were varying from 10 to 1760. Then, the numbers $\mathcal{H}_P(\mathcal{G}_n, k)$ and $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ were computed. Figure 7 captures the results of the experiment. Each plot contains two curves. The first curve represents the changes in average $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ for forty randomly generated graphs of equal number of edges. The second curve represents the variation of the average $\mathcal{H}_P(\mathcal{G}_n, k)$ for the same forty graphs. The trends of these two curves are very similar for different values of k.

Table IV contains the correlation between $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ and $\mathcal{H}_P(\mathcal{G}_n, k)$ calculated for the 560 randomly generated graphs for each value of k.

 ISBN: 978-1-61208-242-4

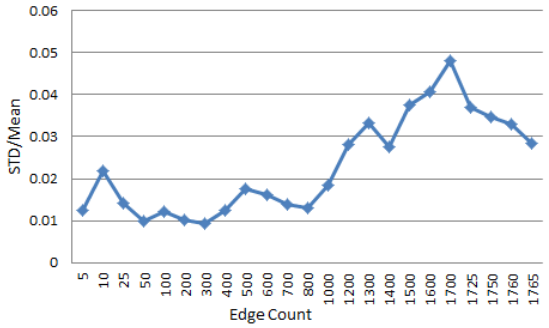


Figure 5. Standard deviation vs. average of the $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ for a number of different permutations of nodes for the same graph. The horizontal axis is labelled with the number of edges of the graph.

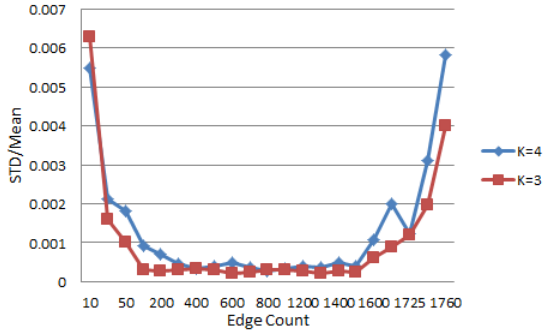


Figure 6. Standard deviation vs. average of the $\mathcal{H}_P(\mathcal{G}_n, k)$ of a number of different permutations of nodes for the same graph. The horizontal axis is labelled with the number of edges of the graph. Each curve corresponds to one value of k.

IV. Frequent Items Sets and Compression Ratio

A market basket data set consists of a multiset T of *transactions*. Each transaction t is a subset of a set of items $I = \{i_1, \ldots, i_N\}$. A transaction is described by its characteristic N-tuple $t = (t_1, \ldots, t_N)$, where

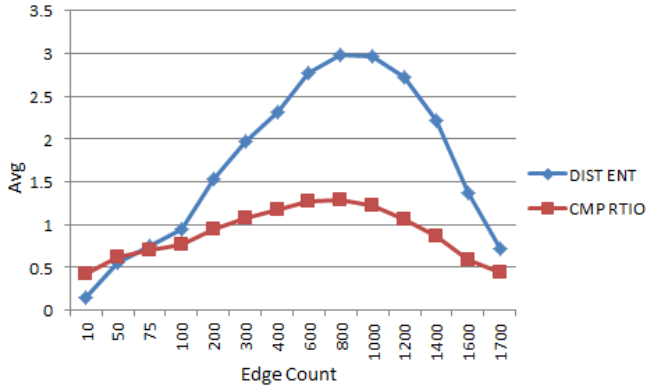
$$t_k = \begin{cases} 1 & \text{if } i_k \in t. \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq k \leq N$. The length of a transaction t is $|t| = \sum_{k=0}^{N} t_k$, while the average size of transactions is $\frac{\sum\{|t| \mid t \text{ in } T\}}{|T|}$.

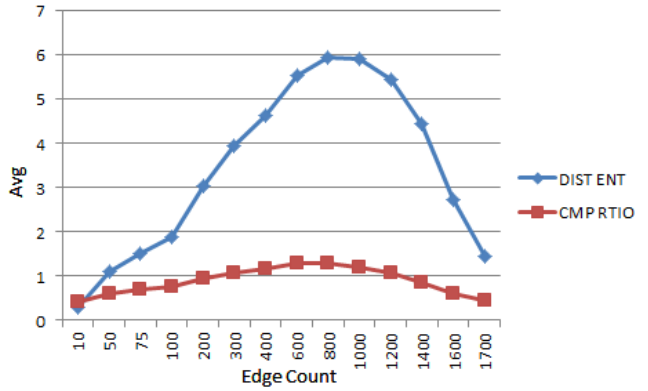
The support of a set of items K of the data set T is the number $\mathsf{supp}(K) = \frac{|\{t \in T \mid K \subseteq t|}{|T|}$. The set of items K is s-frequent if $\mathsf{supp}(K) > s$.

The study of market basket data sets is concerned with the identification of association rules. A pair of item sets (X, Y) is an association rule. Its support, $\mathsf{supp}(X \rightarrow Y)$ equals $\mathsf{supp}(X)$ and its confidence $\mathsf{conf}(X \rightarrow Y)$ is defined as

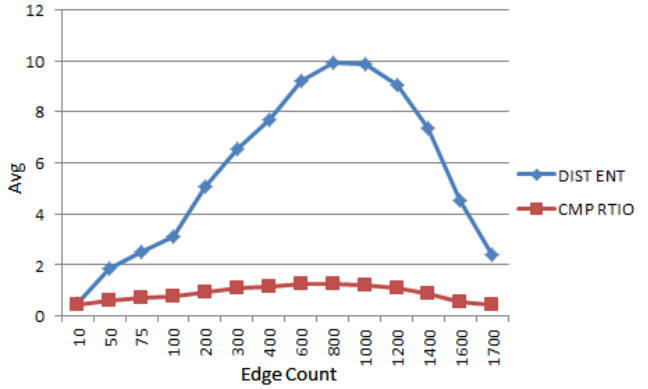
$$\mathsf{conf}(X \rightarrow Y) = \frac{\mathsf{supp}(XY)}{\mathsf{supp}(X)}.$$



$n = 60$ and $k = 3$



$n = 60$ and $k = 4$



$n = 60$ and $k = 5$

Figure 7. Plots of average $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ (CMP RTIO) and average $\mathcal{H}_P(\mathcal{G}_n, k)$ (DIST ENT) for randomly generated graphs $\mathcal{G}_n$ of equal number of edges with respect to the number of edges.

 ISBN: 978-1-61208-242-4

Table IV
CORRELATIONS BETWEEN $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ AND $\mathcal{H}_P(\mathcal{G}_n, k)$

k	Correlation
3	0.92073175
4	0.920952812
5	0.919256573

Using the artificial transaction ARMiner generator described in [3], we created a basket data set. Transactions are represented by sequences of bits $(t_1, \cdots, t_N)$. The multiset of M transactions was represented as a binary string of length MN obtained by concatenating the strings that represent transactions.

We generated files with 1000 transactions, with 100 items available in the basket, adding up to 100K bits.

For data sets having the same number of items and transactions, the efficiency of the compression increases when the number of patterns is lower (causing more repetitions). In an experiment with an average size of a frequent item set equal to 10, the average size of a transaction equal to 15, and the number of frequent item sets varying in the set $\{5, 10, 20, 30, 50, 75, 100, 200, 500, 1000\}$, the compression ratio had a significant variation ranging between 0.20 and 0.75, as shown in Table V. The correlation between the number of patterns and CR was 0.544. Although the frequency of 1s and baseline compression ratio were roughly constant (at 0.75), the number of patterns and compression ratio were correlated.

Table V
NUMBER OF ASSOCIATION RULES AT 0.05 SUPPORT LEVEL AND 0.9 CONFIDENCE

Number of Patterns	Frequency of 1s	Baseline compression	Compression ratio	Number of assoc. rules
5	16%	0.75	0.20	9,128,841
10	17%	0.73	0.34	4,539,650
20	17%	0.73	0.52	2,233,049
30	17%	0.76	0.58	106,378
50	19%	0.75	0.65	2,910,071
75	18%	0.75	0.67	289,987
100	18%	0.75	0.67	378,455
200	18%	0.75	0.70	163
500	18%	0.75	0.735	51
1000	18%	0.75	0.75	3

Further, there was a strong negative correlation (-0.92) between the compression ratio and the number of association rules indicating that market basket data sets that satisfy many association rules are very compressible

V. CONCLUDING REMARKS

Compression ratio of a file can be computed fast and easy, and in many cases offers a cheap way of predicting the existence of embedded patterns in data. Thus, it becomes possible to obtain an approximative estimation of the usefulness of an in-depth exploration of a data set using more sophisticated and expensive algorithms. The use of compression as a measure of minability is illustrated on a variety of paradigms: graph data, market basket data, etc. Recent investigations show that identifying compressible areas of human DNA is a useful tool for detecting areas where the gene replication mechanisms are disturbed (a phenomenon that occurs in certain genetically based diseases.

REFERENCES

[1] B. J. L. Campana and E. J. Keogh. A compression based distance measure for texture. In *SDM*, pages 850–861, 2010.

[2] R. Cilibrasi and P. M. B. Vitnyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51:1523–1545, 2005.

[3] L. Cristofor. ARMiner project, 2000.

[4] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proc. 10th ACM SIGKDD Intnl Conf. Knowledge Discovery and Data Mining*, pages 206–215. ACM Press, 2004.

[5] E. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S. Lee, and J. Handley. Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, 14:99–129, 2007.

[6] E. J. Keogh, L. Keogh, and J. Handley. Compression-based data mining. In *Encyclopedia of Data Warehousing and Mining*, pages 278–285. 2009.

[7] H. Mannila. Theoretical frameworks for data mining. *SIGKDD Exploration*, 1:30–32, 2000.

[8] L. Wei, J. Handley, N. Martin, T. Sun, and E. J. Keogh. Clustering workflow requirements using compression dissimilarity measure. In *ICDM Workshops*, pages 50–54, 2006.

[9] T. Welch. A technique for high performance data compression. *IEEE Computer*, 17:8–19, 1984.

 ISBN: 978-1-61208-242-4

A New Measure of Rule Importance Using Hellinger Divergence

Chang-Hwan Lee
Department of Information and Communications
Dongguk University
Seoul, Korea
Email: chlee@dgu.ac.kr

***Abstract*—Many rule induction algorithms generate a large number of rules in data mining problem, which makes it difficult for the user to analyze them. Thus, it is important to establish some numerical importance measure for rules, which can help users to sort the discovered rules. In this paper, we propose a new rule importance measure, called HD measure, using information theory. A number of properties of the new measure are analyzed.**

***Keywords*-Rule Importance; Association; Information Theory;**

I. INTRODUCTION

Determining the importance of rules is an important data mining problem since many data mining algorithms produce enormous amounts of rules, making it difficult for the user to analyze them manually.

The large number of rules generated by the algorithms commonly used makes it impossible for users to take all the rules into consideration. A common way of reducing the number of rules is to pre-filter the output of data mining algorithms according to importance measures. By selecting a subset of important rules out of a larger set of ones, users can focus on what should be of interest to them. Therefore, importance measures of rules play a major role within a data mining process.

Thus, it is important to establish some numerical importance measure for rules, which can help users to sort the discovered rules. However, the choice of a measure responding to a user's needs is not easy. Therefore there is no optimal measure, and a way of solving this problem is to try to find good compromises.

There are two kinds of rule importance measures: the subjective ones and the objective ones. Subjective measures take into account the user's domain knowledge [1] [2], whereas objective measures are calculated using only the data [3] [4] [5] [6]. We are interested in objective measures, and this article focuses on the objective aspect of rule importance.

Numerous measures are used for performance evaluation in machine learning and data mining. In classification learning, the most frequently used measure is classification accuracy while other measures include precision and recall in information retrieval. With new tasks being introduced in knowledge discovery, new measures need to be defined.

Among the objective measures of rule importance, the information-theoretic measures are important and useful since they are based on theoretical background. In addition, there is an interesting parallel to draw between the use of information theory to evaluate rules [7]. As for rule, the relation is interesting when the antecedent provides a great deal of information about the consequent. The information-theoretic measures commonly used to evaluate rule importance are the Shannon conditional entropy [8], the Theil uncertainty coefficient [5], the J-measure [7], and the Gini index [3].

The Shannon conditional entropy measures the average amount of information of the rule given that the condition is true [8]. The Theil uncertainty coefficient measures the entropy decrease rate of the consequent due to the antecedent [5]. The J-measure is the part of the average mutual information relative to the truth of the antecedent [7]. Finally, the Gini index is the quadratic entropy decrease [3]. Even if these measures are commonly used to evaluate association rules, they are all better suited to evaluate classification rules.

In this paper, we propose a new measure of rule importance, called HD measure, based on information theory. We employ Hellinger divergence as a tool for calculating the importance of rule. A number of properties of the new measure are analyzed. The proposed HD measure shows a number of important and necessary properties.

II. HD MEASURE

For the purposes of this paper, a rule is a knowledge representation of the form b $\rightarrow$ a. We restrict the right-hand expression to being a single value assignment expression while the left-hand side may be a conjunction of such expressions.

The basic idea of rule importance starts with the assumption that the value assignments in the left hand side of each rule affects the probability distribution of the right-hand side(target attribute). The target attribute forms its a priori probabilities without presence of any left-hand conditions. It normally represents the class frequencies of the target attribute. However, its probability distribution changes when it is measured under certain conditions usually given as value assignments of other attributes. Therefore, it is a natural definition, in this paper, that the significance of a

 ISBN: 978-1-61208-242-4

rule is interpreted as the degree of dissimilarity between a priori probability distribution and a posteriori probability distribution of the target attribute. The critical part now is how to define or select a proper measure which can correctly measure the instantaneous information.

In this paper, we employ Hellinger divergence as the measure of instantaneous information. The Hellinger divergence was originally introduced by Beran [9], and, in this paper, we modified it in order to use it as the information content of rules. The original Hellinger divergence of variable A given the value of b is defined as

$$\left(\sum_i \left(\sqrt{p(a_i)} - \sqrt{p(a_i|b)} \right)^2 \right)^{1/2} \quad (1)$$

where a_i denotes the value of variable A. It becomes zero if and only if both a priori and a posteriori distributions are identical, and ranges from 0 to 1. In other words, the Hellinger measure is continuous on every possible combination of a priori and a posteriori values. It can be interpreted as a distance measure where distance corresponds to the amount of divergence between a priori and a posteriori distribution. Therefore, we employ Hellinger measure as a measure of divergence, which will be used as the information amount of rules.

In terms of the probabilistic rules, let us interpret the event $A = a$ as the target concept to be learned and the event (possibly conjunctive) $B = b$ as the hypothesis describing this concept. In this paper, we slightly modify the Hellinger divergence. The information content of a rule (denoted as $IC(b \to a)$) using Hellinger divergence is defined as

$$\begin{aligned} IC(b \to a) &= \left(\sqrt{p(a|b)} - \sqrt{p(a)} \right)^2 + \\ &\quad \left(\sqrt{p(\neg a|b)} - \sqrt{p(\neg a)} \right)^2 \\ &= \left(\sqrt{p(a|b)} - \sqrt{p(a)} \right)^2 + \\ &\quad \left(\sqrt{1 - p(a|b)} - \sqrt{1 - p(a)} \right)^2 \quad (2) \end{aligned}$$

where $p(a|b)$ means the conditional probability of $A = a$ under the condition $B = b$. Notice that Equation (2) has a different form of definition from that of Equation (1). In rule induction, one particular value of the target attribute appears in the right hand side of the pattern, and thus the probabilities for all other values are included in $1 - p(a)$.

In addition, we squared the original form of Hellinger measure because (1) by squaring the original form of Hellinger measure, we could derive a boundary of the Hellinger measure, which allows us to reduce drastically the search space of possible rule rules. (2) the relative information content of each pattern is not affected by the modified Hellinger measure, and (3) the weights between two terms of Hellinger measure provides more reasonable trade-off in terms of their value range.

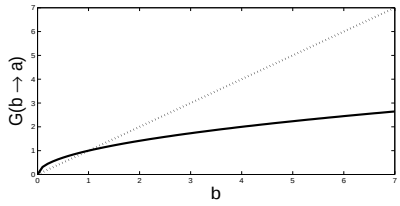


Figure 1. Plot of $\sqrt{p(b)}$ and $p(b)$

Another criteria we have to consider is the *generality* of the rules. The basic idea behind generality is that the more often left-hand side occurs for a rule, the more useful the pattern becomes. The left-hand side must occur relatively often for a pattern to be deemed useful. In this paper, we use

$$G(b \to a) = \sqrt{p(b)} \quad (3)$$

to represent the probability that the rule will occur and, as such, can be interpreted as the measure of rule generality.

The reason for using the square root form of the original probability is that the square root value can represent the generality of events more correctly. The generality of an event(b) increases rapidly when the event first appears. After that, its importance grows slowly when the event has already happened more than enough. Figure 1 compares the plot of $\sqrt{p(b)}$ with that of straight line, which represents $p(b)$.

As shown by the square root function in Figure 1, the value grows rapidly in early state and, then the observation of the event become less important after the event happens a lot. Meanwhile, the linear function of generality, denoted as $p(a)$, grows proportional to the number of events, which does not match with the characteristics of the generality in real world. Another advantage of using the square root form is that we could also derive some boundaries of H measure, described in Property 8 and 9 of the following section.

As a result, by multiplying the generality ($G(b \to a)$) with the information content ($IC(b \to a)$) of the rule, the importance of rule, denoted as $HD(b \to a)$, is given as the following term

$$\begin{aligned} HD(b \to a) &= G(b \to a) \cdot IC(b \to a) \\ &= \sqrt{p(b)} \left[\left(\sqrt{p(a|b)} - \sqrt{p(a)} \right)^2 + \right. \\ &\quad \left. \left(\sqrt{1 - p(a|b)} - \sqrt{1 - p(a)} \right)^2 \right] \quad (4) \\ &= 2\sqrt{p(b)} \left[1 - \sqrt{p(a)p(a|b)} - \right. \\ &\quad \left. \sqrt{(1 - p(a))(1 - p(a|b))} \right] \quad (5) \\ &= 2 \left[\sqrt{p(b)} - \sqrt{p(a)p(ab)} - \right. \\ &\quad \left. \sqrt{(1 - p(a))(p(b) - p(ab))} \right] \quad (6) \end{aligned}$$

which possesses a direct interpretation as a multiplicative measure of the generality and information content of a given

rule.

III. Properties of HD Measure

This section describes the properties of the proposed measure in this paper. Assuming we have such a rule as $b \rightarrow a$, the proposed HD measure has the following properties.

Property 1 : $HD(b \rightarrow a) \geq 0$.

The proof of this property is trivial from the definition of the HD measure given in Equation (4). This property is one of the fundamental properties of rule importance measure since negative importance simply does not make sense in rule mining.

Property 2 : *If a and b are independent, then* $HD(b \rightarrow a) = 0$.

If values a and b are independent with each other, it is known that $p(ab) = p(a)p(b)$. Therefore, from Equation (4), it is clear that $HD(b \rightarrow a) = 0$. In case antecedent attribute and consequent attribute are independent, the resulting importance of the rule ought to be zero. In this sense, this property is an important property.

Property 3 : $HD(b \rightarrow a) \neq HD(a \rightarrow b)$.

With respect to the information content of each rule, $IC(b \rightarrow a) = IC(a \rightarrow b)$. However, $G(b \rightarrow a) \neq G(a \rightarrow b)$. Therefore, $HD(b \rightarrow a) \neq HD(a \rightarrow b)$. Rule $b \rightarrow a$ means there is cause-result relationship between b and a, respectively. This rule does not necessarily mean $a \rightarrow b$.

Property 4 : *Suppose the values of $p(a)$ and $p(b)$ are fixed. When the value of $p(ab)$ increases, the HD measure behaves as follows*

$$HD(b \rightarrow a) = \begin{cases} \searrow & \text{if } p(ab) < p(a)p(b) \\ 0 & \text{if } p(ab) = p(a)p(b) \\ \nearrow & \text{otherwise} \end{cases}$$

The $\searrow$ and $\nearrow$ symbols represent the value of HD measure monotonically increase and monotonically decreases, respectively. From Equation (4),

$$\begin{aligned} \frac{\partial HD(b \rightarrow a)}{\partial p(ab)} &= -2\sqrt{p(a)}\left(\frac{1}{2}\right)\left(\frac{1}{\sqrt{p(ab)}}\right) - \\ &\quad 2\sqrt{1-p(a)}\left(\frac{-1}{2}\right)\left(\frac{1}{\sqrt{p(b)-p(ab)}}\right) \\ &= \sqrt{\frac{1-p(a)}{p(b)-p(ab)}} - \sqrt{\frac{p(a)}{p(ab)}} \end{aligned} \quad (7)$$

Suppose

$$D = \frac{1-p(a)}{p(b)-p(ab))} - \frac{p(a)}{p(ab)} = \frac{p(ab)-p(a)p(b)}{((p(b)-p(ab))p(ab)}$$

(i) If $p(ab) < p(a)p(b)$, then $D < 0$. Therefore, $\frac{\partial HD(b \rightarrow a)}{\partial p(ab)} < 0$.

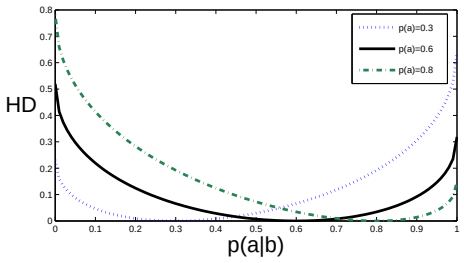


Figure 2. HD values by changing $p(a|b)$

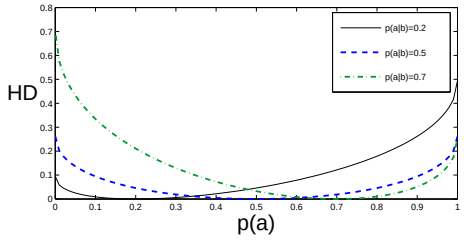


Figure 3. HD values by changing $p(a)$

(ii) If $p(ab) = p(a)p(b)$, then $D = 0$. Therefore, $HD(b \rightarrow a) = 0$.

(iii) If $p(ab) > p(a)p(b)$, then $D > 0$. Therefore, $\frac{\partial HD(b \rightarrow a)}{\partial p(ab)} > 0$. Q.E.D.

This property shows an important characteristic of the HD measure. The HD measure monotonically increases as the degree of deviation from independence between variable of a and b increases.

Property 5 : *Suppose the values of p(a) and p(b) are fixed. When the value of $p(a|b)$ increases, the HD-measure behaves as follows*

$$HD(b \rightarrow a) = \begin{cases} \searrow & \text{if } p(a|b) < p(a) \\ 0 & \text{if } p(a|b) = p(a) \\ \nearrow & \text{otherwise} \end{cases}$$

The proof of this property is straightforward since we get the same results by dividing the probabilities in Property 4 by $p(b)$. Figure 2 show the relationship between HD measure and $p(a|b)$. For simplicity, in Figure 2, the value of $p(b)$ is given as 0.5. The probability $p(a|b)$ can be interpreted as the accuracy of the rule.

Property 6 : *Suppose the values of $p(a|b)$ and $p(b)$ are fixed. When the value of $p(a)$ increases, $HD(b \rightarrow a)$ behaves as follows*

$$HD(b \rightarrow a) = \begin{cases} \searrow & \text{if } p(a) < p(a|b) \\ 0 & \text{if } p(a) = p(a|b) \\ \nearrow & \text{otherwise} \end{cases}$$

Figure 3 show the relationship between HD value and $p(a)$. For simplicity, in Figure 3, the value of $p(b)$ is given as 0.2.

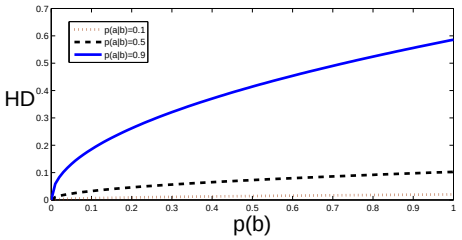


Figure 4. HD values by changing $p(b)$

Property 7 : *HD increases monotonically as the value of p(b) increases.*

This property is true based on Equation 5. Figure 4 shows the relationship between the HD values and $p(b)$ values.

Figure 4 show the relationship between HD value and $p(b)$. For simplicity, in Figure 4, the value of $p(a)$ is given as 0.2.

Property 8 : *In case we add some additional conditions in the rule such as $b \wedge c \rightarrow a$ where C means a set of value assignments. The HD measure of this rule, denoted as HD_2, is bounded as follows.*

$$HD_2 \leq \max\{\sqrt{p(a|b)}\sqrt{p(b)}\left[2\sqrt{m}-2\sqrt{p(a)}\right], \\ 2\sqrt{p(b)}-\sqrt{1-p(a|b)}\sqrt{p(b)}\left[2\sqrt{p(a)}+2\sqrt{1-p(a)}\right]\}$$

where m represents the number of class in the target variable.

With this property, we are able to estimate the boundary of HD measure value without knowing any information about c. Using Property 8, we can predict in advance whether adding conditions in current rule can increase the HD measure. This property is very useful when we generate rules using the proposed HD measure since we can significantly reduce the search space of rule generation.

Property 9 : *Suppose the HD measure of $b \rightarrow a$ and $b \wedge c \rightarrow a$ are HD_1 and HD_2, respectively. In case $p(a|b) = 1$, then $HD_1 \geq HD_2$.*

From $p(b) = p(ab) + p(\neg ab)$ and $p(a|b) = \frac{p(ab)}{p(b)} = 1$,

$$p(\neg ab) = p(b) - p(ab) = 0$$

Therefore,

$$p(a|bc) = \frac{p(abc)}{p(bc)} = \frac{p(abc)}{p(abc)+p(\neg abc)} \\ = \frac{p(abc)}{p(abc)+p(c|\neg ab)p(\neg ab)} = 1 \qquad (8)$$

From Equation (5) and $p(a|b) = 1$,

$$HD_1 = \sqrt{p(b)}\left(2-2\sqrt{p(a)}\right)$$

From Equation (5) and (8),

$$HD_2 = \sqrt{p(bc)}\left(2-2\sqrt{p(a)}\right)$$

Since $p(bc) \leq p(b)$, $\quad HD_2 \leq HD_1$. Q.E.D.

This property is also useful when we generate rules using the proposed HD measure. Like Property 8, we can significantly reduce the search space of rule generation using Property 9.

IV. Conclusions and Future Work

In this paper, we proposed a new information theoretic measure of rule importance, called HD measure. Specifically, we employed Hellinger divergence as the measure of information content of rules, and combined it with the generality of rule. The proposed rule importance show a number of important and interesting characteristics.

The future work of this paper is as follows.

- More analysis of the characteristics of the HD measure
- Apply the HD measure in a rule generation algorithms using real datasets.
- Use the measure as a tool for classification learning.
- Compare the classification performance with current importance measures.

V. Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) (Grant number: 2011- 0023296).

References

[1] B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing the subjective interestingness of association rules," *IEEE Intelligent Systems*, vol. 15, no. 5, pp. 47–55, 2000.

[2] B. Padmanabhan and A. Tuzhilin, "Unexpectedness as a measure of interestingness in knowledge discovery," *Decision Support Systems*, vol. 27, no. 3, pp. 303–318, 1999.

[3] R. J. Bayardo and R. Agrawal, "Mining the most interesting rules," in *KDD*, 1999, pp. 145–154.

[4] X.-H. Huynh, F. Guillet, and H. Briand, "Arqat: An exploratory analysis tool for interestingness measures," in *the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, 334-344, p. 2005.

[5] V. K. P.-N. Tan and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.

[6] P. L. B. Vaillant and S. Lallich, "A clustering of interestingness measures," in *Proceedings of the 7th International Conference on Discovery Science*, 2004, pp. 290–297.

[7] P. Smyth and R. M. Goodman, "An information theoretic approach to rule induction from databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 4, pp. 301–316, 1992.

[8] P. Clark and T. Niblett, "The cn2 induction algorithm," *Machine Learning*, vol. 3, no. 4, pp. 261–283, 1989.

[9] R. J. Beran, "Minimum hellinger distances for parametric models," *Ann. Statistics*, vol. 5, pp. 445–463, 1977.

 ISBN: 978-1-61208-242-4

An Architecture for Semantically Enriched Data Stream Mining

Andreas Textor, Fabian Meyer, Marcus Thoss,
Jan Schaefer, Reinhold Kroeger
Distributed Systems Lab
RheinMain University of Applied Sciences
Unter den Eichen 5, D-65195 Wiesbaden, Germany
{firstname.lastname}@hs-rm.de

Michael Frey
Humboldt-Universität zu Berlin
Department of Computer Science
Rudower Chaussee 25, D-12489 Berlin, Germany
{lastname}@informatik.hu-berlin.de

***Abstract*—Data stream mining (DSM) techniques can be used to extract knowledge from continuous data streams. In this paper, we present an approach providing a modelling and execution architecture for networks of DSM operators. It can be applied in resource-constrained environments, and provides capabilities for semantic enrichment of data streams. This allows processing of streams not only based on information contained in the streams, but also on their semantic contexts. The approach consists of a DSM runtime system, a concept for semantic tagging of stream elements, the integration of semantic information stores, and a domain-specific DSM network description language. A small ambient assisted living scenario is presented as an example application.**

***Keywords*-data stream mining, complex event processing, ontologies, semantic tagging, stream query language, IT management, ambient assisted living.**

I. MOTIVATION

The number of applications where data must be processed in real-time is constantly increasing. Classic data mining approaches are applicable to cases where all data sets are statically accessible in a persistent database, and make use of mathematical and statistical methods to recognize trends, correlations between values, or clustering. If data must be evaluated on the fly, because the amount of data is too high to be stored completely, or because it is inherently continuous, so-called data stream mining (DSM) techniques need to be employed. The goal of data stream mining is the extraction of knowledge from continuous streams of data, e.g., packet streams, monitoring data from sensor networks, streams of log records in applications etc. Usually, knowledge discovery and machine learning approaches are used to achieve this. While data mining usually operates directly on the complete, stored data, DSM tends to make use of supporting application models and data models because of the transient nature of the data to be processed.

In a related field, Complex Event Processing (CEP), streams of distinct simple events are analyzed and events are correlated in order to extract more abstract (complex) events, which may then be inserted into a stream. Any logical or physical change in the observed system is counted as an event, and the events must be processed in a limited time frame, without access to the entirety of data. CEP systems often work using pattern recognition, and abstraction and modeling of events and relationships between events. The processing technology underlying a DSM architecture will often have CEP characteristics as well.

Different approaches for implementations of both DSM and CEP exist, including runtime systems for the extraction and correlation of data and specialized stream query languages. However, few approaches are suitable for the application in resource constrained environments, e.g., home routers with little memory (which becomes relevant in the context of ambient assisted living (AAL) environments). In such environments, a DSM system additionally needs to be highly reconfigurable, as the environment can change quickly and the deployment of new software versions can be non-trivial. On the other hand, AAL installations often provide only a small number of sensors and low frequencies of sensor events. The image that can thus be created from reality tends to be fragmentary, so another motivation for supplementing sensor stream data with additional knowledge is to improve the accuracy of the interpretation of scarce sensor values.

Furthermore, a problem that arises with the application of DSM systems that are integrated into existing application environments is the preservation of semantics of the processed data. Processing needs to be possible depending on semantic information, e.g.: which type of sensor is it, in which room is the sensor located, which are adjacent rooms, used by whom, when, and so on. Equivalent semantic information is necessary in other domains as well.

In this paper, we present an approach for a DSM architecture that satisfies some basic requirements - being able to run in a resource restricted environment, providing facilities for semantic enrichment and processing of data streams that can be expressed in a high-level abstract notation, and dynamic reconfiguration capabilities. The architecture consists of a modular runtime system, which can be adapted to different environments using pluggable input and output connectors, a semantic information store component and a description language for DSM networks.

It is not the goal of the DSM network to replace existing systems, such as the free CEP system Esper [1]. Although

 ISBN: 978-1-61208-242-4

the DSM network can be used stand-alone, it is also possible to combine it with Esper, e.g., for pre-filtering streams. For certain applications (such as processing events in IT management systems), the rate of data in a stream may be much higher than in others (such as processing sensor data in AAL systems), which can make this a viable option.

The paper is structured as follows: Section II describes existing approaches for DSM and CEP systems with regards to the aforementioned requirements, and Section III gives an overview of the proposed architecture, including the structure (III-A), dynamic scripting capabilities (III-B), the semantic information store (III-C), DSM network operators (III-D) and the DSM network description language (III-E). In Section IV, details about the prototypical implementation are given and in Section V an exemplary use case is described. The paper closes with a summary in Section VI.

II. Related Work

An early, popular and widely cited DSM framework was created by the Aurora [2] project. It was designed as a single-site solution focusing on centralised high performance stream processing. Subsequent projects extended the Aurora architecture for distributed processing, notably Medusa [3], also providing a hardware platform for the distributed nodes, and Borealis [4], which resulted from a joint effort of the Aurora and Medusa teams, aiming at a better integration of modelling and distribution aspects. These architectures employ query languages that are syntactically and idiomatically derived from SQL and the underlying database query paradigm, like CQL [5], the query language of STREAM [6], another early single-site stream processing engine. In PIPES [7], a processing network similar to our approach is used, which is expressed formally using a generic operator algebra. From an application perspective, PIPES also maps CQL-related query expressions onto operator networks, focusing on query optimisation.

Since the major academic DSM projects closed down years ago after funding had ceased or key members left the projects, research results have been picked up by larger commercial data mining systems from companies like IBM, Oracle and Tibco. Research architectures were commercialised, notably Aurora, which evolved into the StreamBase CEP product [8] and PIPES, which was branded RTM Analyzer [9] and is now part of the Software AG portfolio. The remaining most prominent open and freely available stream processing framework, Esper [1], is a commercial product as well. StreamBase CEP, RTM Analyzer and Esper queries adhere to the SQL paradigm. Regarding the categories established in [10], they would be classified as using "Data Stream Query Languages", whereas the internal PIPES representation and our approach use "Composition-Operator-Based Event Query Languages".

A more generalised view on event processing and applications relating to data stream mining problems has been stated by a group of IBM researchers in [11]. Especially Processing can be linked to repositories possibly providing contextual knowledge, and events can be "enriched" with additional information, a mechanism we refer to as "tagging". Still, the main goal presented is to detect the occurrence of complex events for immediate reaction, while we strive to interrelate event stream information with long-term semantic model information.

In [12], semantic enrichment of events and subsequent complex event processing using semantic rules and ontologies is described. An architectural vision is given that resembles our concept of mapping events to elements of an ontology through semantic tagging and rule processing. The resulting complex events are not fed back to the ontology, though, but meant to be delivered to the user.

Concerning application orientation related to our work, the DigiHome [13] platform must be mentioned as a CEP-based control architecture targeting AAL environments. It does not rely on an explicit semantic model, but it integrates a large number of real-world devices and abstractions from smart living environments, managing home automation tasks through event stream processing within a service-oriented architecture. The DogOnt [14] modeling language for Intelligent Domotic Environments provides an ontology which allows to formalize all the aspects of the environment and a rule language which can automatically generate proper states and functionalities for domotic devices.

III. Architecture

The architecture presented here adds modelling and scripting facilities to a set of core components and operators, all of which are discussed briefly in the following sections.

A. Data Stream Mining Network

At the heart of our architecture, shown in Figure 1, data stream processing takes place by feeding data into a processing network, which analyses and manipulates stream contents and provides analysis results in the form of output streams or appropriate signalling of results to a monitoring, reporting or result storage environment. The basic building blocks of a DSM network, and thus also the main elements of the framework presented here, are operators manipulating data streams and queues transporting data between the operators.

Our queues are instances of unidirectional FIFO transport lanes for data stream elements ("packets") with basically unlimited storage capacity and one dedicated end for input and output of packets, respectively. Operators realise a given data stream processing functionality ("operator type") that also defines the number of dedicated input and output connection points for queues. Access to the queues is operator-initiated, that is, operators actively push output data into queues and pull input data from queues linked to them. A

 ISBN: 978-1-61208-242-4

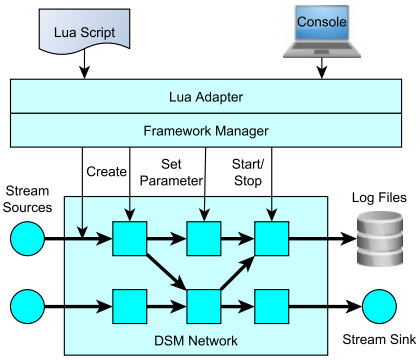


Figure 1. Core DSM Architecture

queue endpoint may be connected to one operator connection point at most.

This boilerplate definition of an abstract operator is essential for the specification of processing networks, since, differently from query-based specification languages like CQL, the operator/queue structure of the network remains exposed up to the specification level. Especially the consistent view of data as streams by all operator types is different from the query approach, which requires the embedding of conversion operators from streams to relational views and back in order to apply SQL-like queries. The format of messages sent through the queues is a compact binary encoded representation of tuples of free-form name-value pairs whose structure and semantics, except for a timestamp value, are defined at application level.

For the creation of a DSM network, operators and queues are instantiated separately one at a time, and then linked as needed such that the resulting data flows and the processing through the operators realise the overall DSM application. Creation and linkage of the network elements is achieved through utilising factory and control functions of a central control entity, the "framework manager".

Operators are active components, i.e., there is at least one thread of control per operator. Currently, there is no additional control over scheduling parameters to adapt resources to asymmetric distributions of processing load among operators, only the default OS scheduling model is used with native threads. For load shedding, there is no central component like in the Borealis architecture, instead, load shedding operators are used that must be explicitly inserted into mining network paths, similar to the concept introduced by CQL.

B. Dynamic Scripting

For some DSM applications, static configurations of operator networks are sufficient. For many problems though, changes in the queries to be conducted can be expected, i.e., in fault diagnosis scenarios, or changing numbers and content of input data streams, all of which leads to change requests targeting DSM network or operator configuration. Dynamic reconfiguration of both operators and network layout are provided by a scripting interface. It is realised as a thin layer that wraps the framework manager, operator and queue interfaces.

For the scripting language, Lua [15] was chosen, because it is considered a mature, sufficiently popular language with a compact native implementation on all mainstream platforms. Lua is object-oriented, which maps well onto the object-oriented design of the DSM framework implementation. Thus, most operations available at the C++ level can be used in the Lua scripting environment.

Technically, the functionality of the internal mining framework manager is thus externalized to an interactive console, which uses a Lua interpreter to execute scripting commands. The manipulation of network elements is expressed as a Lua statement calling methods of Lua objects immediately delegating the method call to corresponding C++ method implementations. Besides being able to alter the network dynamically and interactively using the console interpreter, the same approach can be used to execute prepared Lua script fragments. Through a script, a complete network configuration can be set up, with the added benefit of having a powerful scripting language to further express external dependencies, network variants or to enhance the compactness, expressiveness, and readability of the configuring actions by using loops and other control structures of the Lua language.

C. Semantic Information Store

The data streams in the mining framework contain information that can vary, depending on the originating source of the data (i.e., different textual or binary representations). Therefore, processing of the streams depending on their semantic context is only possible if this context is provided from the outside. For this task, the semantic information store provides access to semantic information in the form of OWL (Web Ontology Language) ontologies. The store is a separate component that is intended to run on a node with more computational resources, as the ontologies may need more memory than what is available on the data source or data stream processing nodes.

The ontologies in the store consist of two parts: A domain ontology, and, where necessary, automatically updated data values extracted by the DSM network. Domain ontologies explicitly model those aspects of the application subject to the DSM process which are not contained in the processed data streams. This allows the semantic enrichment of streams with either actual context information or with references to context information. For cases in which subsequent processing steps depend on context information, corresponding values can be retrieved from the information store and inserted into the data stream. When the original source of a datum is expected to be removed in further processing steps, then a reference to an ontology entity that describes the data source or the original data type of the date, can suffice.

 ISBN: 978-1-61208-242-4

Likewise, current values can be set as state information in the semantic data store in order to preserve them for logging, aggregation or tracking purposes before they are processed in the DSM network and do not remain in the data stream.

D. Mining Network Operators

The approach of interconnected operators is the central concept of the mining framework architecture. Hence, the user needs to be offered a set of operators which can be configured as an interacting network. Most of the operators presented here are already implemented and integrated into the mining framework. There are four classes in which operators can be divided: Analysis, Structural, Knowledge and System operators.

Analysis operators offer functionality for arithmetic and statistical evaluation of numerical values of bypassing data stream elements. They produce new data streams or enrich existing data streams by new elements that contain the result of the analysis process. The operators can either be configured to analyse elements in sliding time windows or sliding element windows. Supported operators include arithmetics and standard deviation, element counter and average for statistics. Further operators like minimum, maximum etc. could be added easily.

Structural operators do not change the state of data stream elements but affect their routing in the network. Mostly, they observe the attributes of bypassing elements and perform a configured routing action. They can decide to either forward or drop elements based on configured filtering conditions (`filter`), or to split incoming data stream elements to the operator's set of output streams using configurable classification characteristics (`split`). Corresponding to the split operator, there is an operator which handles a set of input streams and merges incoming elements to a single output stream (`join`) and the extension of a single stream with additional elements (`ejoin`).

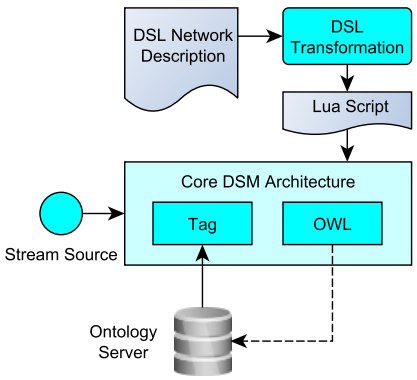


Figure 2. Overall DSM Architecture

Knowledge operators work on semantically enriched data streams. In particular, operations used in different knowledge operations are applied to a stream. This includes the application of OCL (`ocl`) and SWRL rules (`swrl`) to a semantically enriched data stream, like shown in Figure 2.

In order to add semantic tags to a stream in the first place, the tag operator (`tag`) utilises the semantic information store described in Section III-C as an external knowledge base. A SPARQL query is configured for the operator which is executed on the knowledge base for every bypassing element. The query may contain wildcards which are bound to the elements attribute state and returns a result set of key-value-pairs which are added to the element, the so-called "semantic tags". These tags contain information which is normally not present in the data stream but can only be retrieved from the knowledge base. Semantically enriched streams are streams where elements are mapped one-to-one to elements of a knowledge representation. Correspondingly, the untag operator (`untag`) removes semantic enrichments from streams again.

An example for a more complex operator is the decision tree operator (`dtree`). As opposed to the tag operator, it does not have an external knowledge base which already contains information, but builds a classifier using the knowledge contained in the bypassing stream elements. The algorithm used for the implementation is the VFDT (Very Fast Decision Tree learner) described in [16].

System operators offer functionality for loading streams into the DSM system (`in`), output of streams into files (`out`) and on screen (`print`) and logging (`log`). The different possible sources of streams are specified using internationalized resource identifiers (IRI).

E. Mining Network Description Language

The mining network description language (mDSL) is a metamodel-based domain specific language (DSL) for the definition of queries for the DSM system. With it, a domain expert specifies how a stream is structured and what type of sources and sinks for streams exist, what type of operators are executed on a stream, and how they are interconnected. In particular, mappings of model elements of knowledge representations to elements of a stream as well as operations on semantically enriched streams are defined. At the mDSL specification level, streams and operations on a stream are always typed.

Unlike in the DSM network realisation, elements of a stream in mDSL have a unique name and a data type. Streams have unique names as well and consist of one or more elements which are specified within a data structure. At present, integer, floating point numbers, booleans and character strings are supported. In addition, variables can be specified which consist of a unique name, data type and value; they are used within operators. Operators of the mDSL, covering all operator types introduced in the previous section, can be interpreted as functions on streams. Code written in the mDSL can be modularized into packages and re-used through a generic resource import mechanism

 ISBN: 978-1-61208-242-4

that can also be used for loading files provided by different knowledge representations.

mDSL is based on a metamodelling approach. Concepts of the mDSL refer to UML classes and relationships among them. As a concrete syntax of the mDSL, a textual representation was chosen. Semantics are defined using annotations in the mDSL metamodel and in the model-to-text (M2T) transformation.

When application and system design are defined through metamodels, model transformations can be established that generate executable applications from formal software models, thus supporting a model-driven software development (MDSD) process. Specifically, M2T transformations create a textual representation of the input model. Figure 2 depicts the transformation process applied to mDSL -based models. Descriptions of the mining network written in the mDSL are transformed to Lua using an M2T transformation.

Using a metamodel-based approach also facilitated the integration of knowledge representations. Here, the ontology definition metamodel (ODM) [17], a metamodel-based representation of OWL specified by the OMG was used. Using ODM, our concept aims at linking model-driven software development with the use of ontologies. Shared and disjoint concepts among OWL and UML are identified and the knowledge representations are integrated by references in classes and relationships among classes in the mDSL metamodel. Thus, an mDSL-based operator class for mapping elements of an OWL ontology to elements of a stream can reference an OWL class represented as an UML class in the ODM metamodel. This metamodelling approach is very flexible since it allows to add further metamodel-based knowledge representations to the language. A limiting factor are the methods and concepts of a knowledge representation, since they are shared among all knowledge representations covered by the language.

IV. Implementation

For the architecture presented, a prototypical implementation has been created, and some of its aspects are highlighted in the following sections.

A. Data Stream Mining Network

The framework with all of its core components and operators is natively implemented using C++ and the Boost [18] libraries. For operators, a base class provides both the basic implementation of thread handling and queue connectors as a basis for new, concrete operator implementations. Every operator is implemented within a shared library of its own, such that it can be dynamically loaded and the operator instances can be created at run-time.

The operator thread implementation basically relies on operating system threads, which are encapsulated by the `thread` abstraction of Boost libraries. Of course, the concurrency introduced by active operators incurs race conditions during queue access. Valid access semantics are guaranteed by a combination of measures: First, at any given point in time, a message is owned by exactly one queue or operator, which avoids concurrency issues for access to message contents. A lock-free queue implementation is used in combination with memory barriers to ensure both correctness and minimal locking.

The encoding of message contents is handled by a BSON [19] implementation taken from the MongoDB [20] project. BSON provides a compact binary encoding of structured data. It also offers dynamic protocol buffer handling based on Boost memory object abstractions like smart pointers, which ensures that message memory is freed when a message is no longer referenced by an operator or by a queue.

The native implementation and the compact message format were chosen with resource-constrained environments in mind. The virtual image size of an engine running a network of 30 operators is around 50 MiB, and the message size is almost devoid of overhead beyond the binary representation of field names and values. For the deployment of DSM networks on small appliances like those found in home automation environments, minimisation of the overhead of both memory footprint and computing is essential because in living environments the limits of acceptance for energy consumption and noise and heat emission of computing devices are critical. Performance optimisations concerning throughput and latency have not been carried out aggressively yet, the same applies to memory pooling strategies for messages of similar size. The processing rate for a simple, linear three-operator setup with a numeric threshold filtering operator is in the range of 200k messages per second on a 2 GHz dual core PC system, which is still half a magnitude below the Java-based Esper engine.

B. Knowledge Operators

The assembly for most operators that have been presented in Section III-D is rather simple, but especially the knowledge operators have a complex structure for the integration of external knowledge bases. Hence, the tag operator is examined in this section.

The external knowledge base used by the tag operator is an ontology offered by the semantic information store presented in Section III-C. Since the exported interface of the store is an OSGi service, the interprocess communication proves difficult. Therefore, the semantic information store was extended by a socket-based communication channel that takes a SPARQL query string as input, performs the query on the knowledge base, and writes the resulting tuples of IRIs to the socket. The query is constructed using a template, containing wildcards for stream element specific data, which is configured for each tag operator instance. When an element is taken from the operator's input stream, its attribute values are bound to the wildcards and the query is sent to the semantic information store. The resulting data

 ISBN: 978-1-61208-242-4

tuples are added to the elements attributes and the element is put into the operators output stream, so that subsequent operators can use the semantically enriched data for further knowledge-based operations.

V. APPLICATION

As described earlier, one of the major motivations for the DSM is the application in AAL environments. In this section, it is shown how our DSM architecture can be applied to a small but representative example application from an AAL setting.

The scenario presented here consists of two light sensors, one outside and one inside a house, a sensor to detect a person in the house and an actor to control window blinds. Furthermore, the network contains a decision tree operator to classify sensor values. The goal of the given scenario is to decide, on the basis of the given sensor values, whether or not to open or close the window blinds, and then perform the particular action. In a simple example such as this, a set of rules or a purpose-built operator could be easily employed to make a decision. In more realistic situations, though, more sensors and additional information would have to be taken into account (e.g., more persons, a notion of time, whether the TV is turned on etc.).

```
package example {
  stream sensors {
    int id;
    float value;
  }
  stream actors { int action; }
  stream queryresult { bool return_value; }
  sensors indoors, outside, person, aggregated;
  actors blinds;
  queryresult result;

  indoors = in("http://wwwvs.cs.hs-rm.de/dsm/sensors/
    indoors/#1", "localhost", 9597);
  outside = in("http://wwwvs.cs.hs-rm.de/dsm/sensors/
    outside/#1", "localhost", 9595);
  person = in("http://wwwvs.cs.hs-rm.de/dsm/sensors/
    person/#1", "localhost", 9596);
  aggregated = join(indoors, outside, person);

  result = sparql({SELECT ?"blinds"  WHERE { ?"room"
    <http://wwwvs.cs.hs-rm.de/dsm/contains> ?"blinds" .
    ?"room" <http://wwwvs.cs.hs-rm.de/dsm/contains>
    "indoorslightsensor"}}, result.return_value,
  aggregated);

  blinds = dtree("PERSON" {"IN", "OUT"}, "LIGHT_OUTSIDE"
    {"YES", "NO"}, "LIGHT_INDOORS" {"YES", "NO"}, 0.0001,
    0.025, 0.98, 450, "BLINDS" {"YES", "NO"}, result);
}
```

Listing 1. Example application

Each of the sensors provides a separate data stream in the DSM. The streams from the three sensors are merged into one stream, which in turn is passed through an instance of the tag (“sparql”) operator that enriches the stream with the information about which blinds should be controlled. This is information that is not present in the stream - depending on the room, in which the light sensor is situated, different window blinds could be selected. The enriched stream is then passed to the decision tree operator, which decides to open or close the window blinds, and finally to the actuator.

Listing 1 shows the DSM network description language code for the scenario described above.

VI. SUMMARY AND FUTURE WORK

In this paper, we presented an approach for a data stream mining architecture that takes into account the requirements for being applied in resource-constrained environments and for providing facilities for semantic enrichment and processing of data streams. The approach includes a model for DSM networks, a modular runtime system, a semantic information store component and a description language for DSM networks.

The DSM network is application agnostic, as the framework and the operators operate on data streams which are not further specified, i.e., contain untyped information, except for the internal typing necessary to encode binary data representations. This design decision allows the construction of highly specialised and thus resource saving operators. Nevertheless, a set of standard operators was described and implemented, covering both structural operators such as join and merge, numerical operators such as sum and average and classification operators such as the VFDT operator.

Using the semantic information store, stream data can be enriched with semantic information. The tag operator can be used in the DSM network in places where additional information from the knowledge base external to the actual DSM elements is required for further processing. Semantic information specific to the data source or data stream can be added to the stream, such that subsequent operators (e.g., classification operators) can incorporate the additional information.

The DSM description language is a metamodel-based domain specific language that can be used to specify a network. Programs in the language are compiled into a sequence of commands to set up the DSM network incrementally. While, for performance reasons, the elements of data streams in the runtime system are not explicitly typed, the specification language does support type information about the streams and can use it to ensure consistency of the network.

Next steps in the project include further performance improvements and the implementation of more DSM network operators: While the description language already supports the specification of operators for the query languages OCL and SWRL, this has yet to be implemented in the runtime system. An operator that feeds back semantic tag information from result streams to an ontology will be realised to complete the semantic transfer cycle. It is also planned to apply the approach to a scenario dealing with a larger AAL infrastructure.

 ISBN: 978-1-61208-242-4

REFERENCES

[1] EsperTech, "Esper Complex Event Processing System," http://esper.codehaus.org/. Last access 2012-07-10.

[2] D. Carney, U. Çetintemel, M. Cherniack, C. Convey, S. Lee, G. Seidman, M. Stonebraker, N. Tatbul, and S. Zdonik, "Monitoring Streams: A New Class of Data Management Applications," in *Proceedings of the 28th international conference on Very Large Data Bases*, ser. VLDB '02. VLDB Endowment, 2002, pp. 215–226. [Online]. Available: http://dl.acm.org/citation.cfm?id=1287369.1287389

[3] S. Z. Sbz, S. Zdonik, M. Stonebraker, M. Cherniack, U. C. Etintemel, M. Balazinska, and H. Balakrishnan, "The Aurora and Medusa Projects," *IEEE Data Engineering Bulletin*, vol. 26, 2003.

[4] D. J. Abadi, Y. Ahmad, M. Balazinska, M. Cherniack, J. hyon Hwang, W. Lindner, A. S. Maskey, E. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik, "The Design of the Borealis Stream Processing Engine," in *In CIDR*, 2005, pp. 277–289.

[5] A. Arasu, S. Babu, and J. Widom, "The CQL Continuous Query Language: Semantic Foundations and Query Execution," *The VLDB Journal*, vol. 15, pp. 121–142, June 2006. [Online]. Available: http://dx.doi.org/10.1007/s00778-004-0147-z

[6] A. Arasu, B. Babcock, S. Babu, J. Cieslewicz, M. Datar, K. Ito, R. Motwani, U. Srivastava, and J. Widom, "STREAM: The Stanford Data Stream Management System," Stanford InfoLab, Technical Report 2004-20, 2004. [Online]. Available: http://ilpubs.stanford.edu:8090/641/

[7] J. Krämer, "Continuous Queries over Data Streams - Semantics and Implementation," Ph.D. dissertation, Philipps-Universität Marburg, 2007.

[8] StreamBase Systems, Inc., "StreamBase CEP," http://www.streambase.com. Last access 2012-07-10.

[9] RTM Realtime Monitoring GmbH, "RTM Analyzer," http://www.realtime-monitoring.de. Last access 2012-07-10.

[10] M. Eckert, "Complex Event Processing with XChange EQ: Language Design , Formal Semantics, and Incremental Evaluation for Querying Events," Ph.D. dissertation, Ludwig-Maximilans-Unversität München, 2008.

[11] C. Moxey, M. Edwards, O. Etzion, M. Ibrahim, S. Iyer, H. Lalanne, M. Monze, M. Peters, Y. Rabinovich, G. Sharon, and K. Stewart, "A Conceptual Model for Event Processing Systems, IBM Form Number REDP-4642-00," IBM Redguide publication, 2010. [Online]. Available: http://www.redbooks.ibm.com/abstracts/REDP4642.htm

[12] K. Teymourian and A. Paschke, "Enabling knowledge-based complex event processing," in *Proceedings of the 2010 EDBT/ICDT Workshops*, ser. EDBT '10. New York, NY, USA: ACM, 2010, pp. 37:1–37:7. [Online]. Available: http://doi.acm.org/10.1145/1754239.1754281

[13] D. Romero, G. Hermosillo, A. Taherkordi, R. Nzekwa, R. Rouvoy, and F. Eliassen, "The DigiHome Service-Oriented Platform," *Software: Practice and Experience*, 2011. [Online]. Available: http://hal.inria.fr/inria-00563678

[14] D. Bonino and F. Corno, "DogOnt - Ontology Modeling for Intelligent Domotic Environments"," in *Proceedings of the 7th International Semantic Web Conference*, ser. LNCS. Springer, 2008, pp. 790–803.

[15] R. Ierusalimschy, *Programming in Lua, Second Edition*. Lua.Org, 2006.

[16] P. Domingos and G. Hulten, "Mining High-speed Data Streams," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '00. New York, NY, USA: ACM, 2000, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/347090.347107

[17] Object Management Group, "Ontology Definition Metamodel," http://www.omg.org/spec/ODM/1.0/PDF. Last access 2012-07-10.

[18] B. Dawes, D. Abrahams, and R. Rivera, "Boost C++ Libraries," http://www.boost.org. Last access 2012-07-10.

[19] D. Merriman, "BSON - Binary JSON," http://bsonspec.org.

[20] 10gen, Inc., "mongoDB," http://www.mongodb.org. Last access 2012-07-10.

 ISBN: 978-1-61208-242-4