DATA ANALYTICS 2014

The Third International Conference on Data Analytics

ISBN: 978-1-61208-358-2

August 24 - 28, 2014

Rome, Italy

DATA ANALYTICS 2014 Editors

Fritz Laux, Reutlingen University, Germany

Panos M. Pardalos, University of Florida, USA

Alain Crolotte, Teradata Corporation - El Segundo, USA

DATA ANALYTICS 2014

Forward

The Third International Conference on Data Analytics (DATA ANALYTICS 2014), held on August 24 - 28, 2014 - Rome, Italy, continued the 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 2014 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to the DATA ANALYTICS 2014. 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 2014 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 2014 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 Rome provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

DATA ANALYTICS 2014 Chairs:

DATA ANALYTICS Advisory Chairs
Fritz Laux, Reutlingen University, Germany
Lina Yao, The University of Adelaide, Australia
Eiko Yoneki, University of Cambridge, UK

Takuya Yoshihiro, Wakayama University, Japan
Felix Heine, University of Applied Sciences & Arts Hannover, Germany
Dominik Slezak, University of Warsaw & Infobright Inc., Poland
Panos M. Pardalos, University of Florida, USA
Michele Melchiori, Università degli Studi di Brescia, Italy
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
Sergio Ilarri, University of Zaragoza, Spain
Les Sztandera, Philadelphia University, USA
Prabhat Mahanti, University of New Brunswick, Canada
Dominique Laurent, University of Cergy Pontoise, France
Ryan G. Benton, University of Louisiana at Lafayette, USA
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Stratos Idreos, Harvard University, USA
Andrew Rau-Chaplin, Dalhousie University, Canada
Takuya Yoshihiro, Wakayama University, Japan

DATA ANALYTICS Industry/Research Liaison Chairs
Qiming Chen, HP Labs - Palo Alto, USA
Alain Crolotte, Teradata Corporation - El Segundo, USA
Tiziana Guzzo, National Research Council/Institute of Research on Population and Social Policies - Rome, Italy
Shlomo Geva, Queensland University of Technology - Brisbane, Australia
Farhana Kabir, Intel, USA
Prabhanjan Kambadur, IBM TJ Watson Research Center, USA
Serge Mankovski, CA Technologies, Spain
Sumit Negi, IBM Research, India
Vedran Sabol, Know-Center - Graz, Austria
George Tambouratzis, Institute for Language and Speech Processing - Athena Research Centre, Greece
Yanchang Zhao, RDataMining.com, Australia
Masahiro Terabe, Mitsubishi Research Institute, Inc., Japan
Marina Santini, Santa Anna IT Research Institute AB, Sweden
Mario Zechner, Know-Center, Austria

DATA ANALYTICS Publicity Chairs
Johannes Leveling, Dublin City University, Ireland
Tim Weninger, University of Illinois in Urbana-Champaign, USA
Roberto Zicari, Johann Wolfgang Goethe - University of Frankfurt, Germany
Shandian Zhe, Purdue University, USA
Michael Schaidnagel, Reutlingen University, Germany

DATA ANALYTICS 2014

Committee

DATA ANALYTICS Advisory Chairs

Fritz Laux, Reutlingen University, Germany
Lina Yao, The University of Adelaide, Australia
Eiko Yoneki, University of Cambridge, UK
Takuya Yoshihiro, Wakayama University, Japan
Felix Heine, University of Applied Sciences & Arts Hannover, Germany
Dominik Slezak, University of Warsaw & Infobright Inc., Poland
Panos M. Pardalos, University of Florida, USA
Michele Melchiori, Università degli Studi di Brescia, Italy
Rudolf Berrendorf, Bonn-Rhein-Sieg University, Germany
Kerstin Lemke-Rust, Hochschule Bonn-Rhein-Sieg, Germany
Sergio Ilarri, University of Zaragoza, Spain
Les Sztandera, Philadelphia University, USA
Prabhat Mahanti, University of New Brunswick, Canada
Dominique Laurent, University of Cergy Pontoise, France
Ryan G. Benton, University of Louisiana at Lafayette, USA
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Stratos Idreos, Harvard University, USA
Andrew Rau-Chaplin, Dalhousie University, Canada
Takuya Yoshihiro, Wakayama University, Japan

DATA ANALYTICS Industry/Research Liaison Chairs

Qiming Chen, HP Labs - Palo Alto, USA
Alain Crolotte, Teradata Corporation - El Segundo, USA
Tiziana Guzzo, National Research Council/Institute of Research on Population and Social Policies - Rome, Italy
Shlomo Geva, Queensland University of Technology - Brisbane, Australia
Farhana Kabir, Intel, USA
Prabhanjan Kambadur, IBM TJ Watson Research Center, USA
Serge Mankovski, CA Technologies, Spain
Sumit Negi, IBM Research, India
Vedran Sabol, Know-Center - Graz, Austria
George Tambouratzis, Institute for Language and Speech Processing - Athena Research Centre, Greece
Yanchang Zhao, RDataMining.com, Australia
Masahiro Terabe, Mitsubishi Research Institute, Inc., Japan
Marina Santini, Santa Anna IT Research Institute AB, Sweden
Mario Zechner, Know-Center, Austria

DATA ANALYTICS Publicity Chairs

Johannes Leveling, Dublin City University, Ireland
Tim Weninger, University of Illinois in Urbana-Champaign, USA
Roberto Zicari, Johann Wolfgang Goethe - University of Frankfurt, Germany
Shandian Zhe, Purdue University, USA
Michael Schaidnagel, Reutlingen University, Germany

DATA ANALYTICS 2014 Technical Program Committee

Mohd Helmy Abd Wahab, Universiti Tun Hussein Onn Malaysia, Malaysia
Sayed Abdel-Wahab, Sadat Academy for Management Sciences, Egypt
Rajeev Agrawal, North Carolina A&T State University - Greensboro, USA
Maik Anderka, University of Paderborn, Germany
Fabrizio Angiulli, University of Calabria, Italy
Annalisa Appice, Università degli Studi di Bari Aldo Moro, Italy
Giuliano Armano, University of Cagliari, Italy
Ryan G. Benton, University of Louisiana at Lafayette, USA
Erik Buchmann, Karlsruhe Institute of Technology, Germany
Luca Cagliero, Politecnico di Torino, Italy
Huiping Cao, New Mexico State University, USA
Michelangelo Ceci, University of Bari, Italy
Federica Cena, Università degli Studi di Torino, Italy
Lijun Chang, University of New South Wales, Australia
Qiming Chen, HP Labs - Palo Alto, USA
Been-Chian Chien, National University of Tainan, Taiwan
Silvia Chiusano, Politecnico di Torino, Italy
Alain Crolotte, Teradata Corporation - El Segundo, USA
Bo Dai, Purdue University, U.S.A.
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
Sherif Elfayoumy, University of North Florida, USA
Wai-keung Fung, Robert Gordon University, UK
Matjaz Gams, Jozef Stefan Institute - Ljubljana, Slovenia
Paolo Garza, Dipartimento di Automatica e Informatica Politecnico di Torino, Italy
Shlomo Geva, Queensland University of Technology - Brisbane, Australia
Amer Goneid, American University in Cairo, Egypt
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
Tiziana Guzzo, National Research Council/Institute of Research on Population and Social Policies - Rome, Italy
Michael Hahsler, Southern Methodist University, U.S.A.
Sven Hartmann, TU-Clausthal, Germany
Felix Heine, Hochschule Hannover, Germany
Quang Hoang, Hue University, Vietnam

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

Comparison of Linear Discriminant Functions by K-fold Cross Validation

Shuichi Shinmura
Faculty of Economics, Seikei Univ.
Tokyo, Japan
shinmura@econ.seikei.ac.jp

***Abstract*— To discriminate two classes is essential in the science, technology, and industry. Fisher defined the linear discriminant function (Fisher's LDF) based on the variance-covariance matrices. It was applied for many applications. After Fisher's LDF, several LDFs such as logistic regression and a soft margin support vector machine (S-SVM) are proposed. But, there are serious two problems of the discriminant analysis. First, the numbers of misclassifications (NMs) or error rates by these LDFs may not be correct because these LDFs cannot discriminate cases on the discriminant hyper-plane correctly. Second, these LDFs cannot recognize the linear separable data properly. Only revised optimal LDF by integer programming (Revised IP-OLDF) resolves these problems. In this paper, we compare seven LDFs by 100-fold cross validation using 104 different discriminant models. It is shown that the mean error rates of Revised IP-OLDF are better than other LDFs in the training and validation samples.**

Keywords- Fisher's linear discriminant function; logistic regreesion; soft margin SVM; Revised IP-OLDF; minimum number of misclassifications; k-fold cross validation.

I. INTRODUCTION

To discriminate two classes or objects is essential in the science, technology, and industry. Fisher defined the linear discriminant function (LDF) to maximize the variance ratio (between classes/within class) [2]. If two classes satisfy the Fisher's assumption that two classes belong to the normal distribution such as N_i (x: m_i, Σ_i) i=1, 2 and $\Sigma_1 = \Sigma_2$, the same LDF is formulated by the plug-in rule in (1).

$$\text{Log}\,(N_1(x: m_1, \Sigma_1) / N_2(x: m_2, \Sigma_2)) = 0 \qquad (1)$$

And it is defined by the variance-covariance matrices explicitly in (2).

$$f(\mathbf{x}) = \{\mathbf{x}-(\mathbf{m}_1+\mathbf{m}_2)/2\}'\sum^{-1}(\mathbf{m}_1-\mathbf{m}_2) \qquad (2)$$

x: p-independent variables (p-features).
$\mathbf{m}_1/\mathbf{m}_2$: mean vectors in class1/class2.
$\sum$: pooled variance-covariance matrix.

Statistical software packages adapt this equation, and many useful methods such as the variable selection methods are developed. It was applied for many applications such as the medical diagnosis, genome discrimination, pattern recognition, the rating of stocks and the pass/fail determination of exams score [16] etc.

The discriminant rule is very simple: If $y_i*f(\mathbf{x}_i) > 0$, $\mathbf{x}_i$ is classified to class1/class2 correctly. If $y_i*f(\mathbf{x}_i) < 0$, $\mathbf{x}_i$ is misclassified. This simplicity may hide the following problems:

1) Problem 1: If there are cases on the discriminant hyper -plane ($f(\mathbf{x}_i)=0$), we cannot discriminate these cases correctly. This is the unresolved problem of discriminant analysis. Until now, most statistical user treats that these cases belong to class1 without any reason. Some statisticians explain that this is decided by the probability because statistics is a study, which is based on the probability. But statistical software adopt former rule. And the medical doctors who use the discriminant analysis in the medical diagnosis are surprised and disappointed by the latter explanation. They devote heart and soul to discriminate the patient near by the discriminant hyper-plane.

2) Problem 2: A hard margin SVM (H-SVM) defines the discrimination of linear separable data clearly. But there are few researches about it. First reason is that Fisher's LDF, logistic regression and soft-margin SVM (S-SVM) cannot recognize linear separable data. Second reason is there are no good research data of linear separable data. Ranges of 18 error rates of Fisher's LDF and quadratic discriminant function (QDF) are [2.2%, 16.7%] and [0.8%, 8.5%] by the pass/fail determination of exams scores [18], nevertheless those are linear separable.

These two problems are resolved by IP-OLDF and Revised IP-OLDF [19] [21].

3) Problem 3: The discriminant functions based on the variance-covariance matrices need to compute the inverse matrices. But if some variables are constant, those are not computed. The generalized inverse matrix technique may be expected to resolve this defect. But the serious problem is found in the special case in QDF [18].

In this research, problem 4 is discussed.

4) Problem 4: After Fisher's LDF, many LDFs are proposed. There are few comparisons of these LDFs. In this research, seven LDFs are compared by k-fold cross validation using 104 different discriminant models of four data such as Fisher's iris data [1], Swiss bank note data [3],

 ISBN: 978-1-61208-358-2

Cephalo Pelvic Disproportion (CPD) data [11], and the student data [14].

II. LINEAR DISCRIMINANT FUNCTIONS COMPARED IN THIS RESEARCH

After Fisher's LDF, QDF and the multi-class discrimination using the Maharanobis distance are proposed in the statistical approach. These methods are formulated by the variance-covariance matrices. In this research, only seven LDFs in this chapter are compared by 100-fold cross validation in order to approach problem 4.

A. Logistic Regression

In the medical diagnosis, the discriminant methods are very important and useful. But, real data scarcely satisfy the Fisher's assumption, especially in the epidemiological study. Therefore, the logistic regression in (3) was developed by Framingham sturdy.

$$\text{Log}\,(P_i/(1-P_i)) = b_1x_1+\ldots+b_px_p+b_0 \quad (3)$$

P_1 / P_2: the probability of the normal / ill class.
$(x_1,\ldots,x_p)$ or $\mathbf{x}$: p-features (independent variables) vector.
$(b_1,\ldots,b_p)$ or $\mathbf{b}$: p-discriminant coefficients vector.
b_0: the constant of LDF.
n: the number of cases (n_1/n_2: the normal / ill class)

B. Support Vector Machine

The regression and discriminant analyses are easily approached by the mathematical programming (MP) because MP can find the minimum / maximum (global optimal) value of function. Before SVM, there are many researches of L_p-norm discriminant functions by linear programming (LP). Stam summarized these researches and sorrowed "statistical users rarely use these functions" [22]. Statistical users use SVM because there are many evaluations of SVM by the real data. On the other hand, there are no evaluations of the MP-based discriminant functions before SVM. Vapnik proposed three kinds of SVM, such as H-SVM, S-SVM and kernel SVM [23]. H-SVM in (4) indicates the discrimination of linear separable data definitely. Cases $\mathbf{x}_i$ are classified correctly by the support vectors (SVs). The object function minimizes (1/ the distance between two SVs). This is to maximize the distance between two SVs. It has been proven that the generalization ability of H-SVM is good.

$$\text{MIN} = \|\mathbf{b}\|^2/2; \quad y_i*(\mathbf{x}_i'\mathbf{b}+b_0) \geq 1; \quad (4)$$
$$y_i = 1 / -1 \text{ for } \mathbf{x}_i \in \text{class1/class2}.$$

Real data are rarely linear-separable. Therefore, S-SVM has been defined in (5). S-SVM permits certain cases that are not discriminated by SV ($y_i*f(\mathbf{x}_i) < 1$). The second objective is to minimize the summation of distances of misclassified cases (Σe_i) from SV. These two objects are combined by defining "penalty c." The Markowitz portfolio model to minimize risk and maximize return is as same as S-SVM. However, the return is incorporated in the constraint, and the objective function minimizes only risk. The decision maker chooses a good solution on the efficient frontier. On the contrary, S-SVM does not have a rule to determine c. Nevertheless, it can be solved by an optimization solver. In this research, we try to evaluate two S-SVMs (c =10^4 and 1).

$$\text{MIN} = \|\mathbf{b}\|^2/2 + c*\Sigma e_i; \quad y_i*(\mathbf{x}_i'\mathbf{b}+b_0) \geq 1 - e_i; \quad (5)$$
e_i : non-negative decision variable.
c: penalty c to combine two objectives.

C. Heuristic-OLDF and IP-OLDF

Shinmura and Miyake [10] developed the heuristic algorithm of OLDF based on MNM criterion affected by Warmack and Gonzalez [24]. This OLDF solved only five features (5-vars) model of CDP data because of the lack of the CPU power.

SAS was introduced into Japan in 1978 [6]. LINDO was introduced into Japan in 1983. Several regression models are formulated by MP [8]. Least-squares method can be solved by QP, and Least Absolute Value (LAV) regression is solved by LP. Without a survey of previous research, the formulation of IP-OLDF [12][13] can be defined as in (6).

$$\text{MIN} = \Sigma e_i; \quad y_i*(\mathbf{x}_i'\mathbf{b}+1) >= -M*e_i; \quad (6)$$
e_i: 0/1 integer variable corresponding to $\mathbf{x}_i$.
M: 10,000 (Big M constant).

This notation is defined on p-dimensional coefficients space because the constant of LDF is fixed to 1. In pattern recognition, the constant is a free variable. In this case, the model is defined on (p+1)-coefficients space, and we cannot elicit the same deep knowledge as with IP-OLDF. This difference is very important. IP-OLDF is defined on both p-dimensional data and coefficients spaces. This is very important to find new facts of the discriminant analysis [14]. We can understand new knowledge of the discriminant analysis about the relation between the NMs and LDFs clearly. This relation tells us the following new facts and shows a clue of problem solving.

1) Fact 1: Optimal Convex Polyhedron

The linear equation $H_i(\mathbf{b}) = y_i*(\mathbf{x}_i'\mathbf{b}+1) = 0$ divides p-dimensional coefficients space into plus and minus half-planes ($H_i(\mathbf{b}) > 0$, $H_i(\mathbf{b}) < 0$). If $\mathbf{b}_j$ is in the plus half-plane, $f_j(\mathbf{x}) = y_i*(\mathbf{b}_j'\mathbf{x}+1)$ discriminates $\mathbf{x}_i$ correctly because $f_j(\mathbf{x}_i) = y_i*(\mathbf{b}_j'\mathbf{x}_i+1) = y_i*(\mathbf{x}_i'\mathbf{b}_j+1) > 0$. On the contrary, if $\mathbf{b}_j$ is included in the minus half-plane, $f_j(\mathbf{x})$ cannot discriminate $\mathbf{x}_i$ correctly because $f_j(\mathbf{x}_i) = y_i*(\mathbf{b}_j'\mathbf{x}_i+1) = y_i*(\mathbf{x}_i'\mathbf{b}_j+1) < 0$. The n linear equations $H_i(\mathbf{b})$ divide the coefficients space into a finite number of convex polyhedrons. Each interior point of a convex polyhedron has a unique NM that is equal to the number of minus half-planes of n linear equations. We define the "Optimal Convex Polyhedron (OCP)" as that for which NM is equal to MNM.

2) Fact 2: $\text{MNM}_q \geq \text{MNM}_{(q+1)}$

Let us MNM_q be MNM of q-vars model, and $\text{MNM}_{(q+1)}$ be MNM of (q+1)-vars model adding one variable to the

former. The proof is very easy. The OCP of q-vars model is concluded in (q+1)-discriminant coefficients space. At least, we know there exits the convex polyhedron in (p+1)-coefficients space, NM of which is MNM_q.

3) Fact 3: Two kinds of the discrimination

If MNM_q =0, all MNMs including these q-features are zero. IP-OLDF found Swiss bank note data is linear separable by 2-features such as (X4, X6). It consisted of two kinds of bills: 100 genuine and 100 counterfeit bills. There were six features: X1 was the length of the bill; X2 and X3 were the width of the left and right edges; X4 and X5 were the bottom and top margin widths; X6 was length of the image diagonal. A total of 63 ($=2^6$-1) models were investigated. We had better considered about two types of discriminations: 16 linearly separable discriminant models, and other 47 models. This data is adequate whether or not LDFs can discriminate linearly separable data correctly.

D. Revised IP-OLDF

The Revised IP-OLDF in (7) can find the true MNM because it can directly find the interior point of the OCP. This means there are no cases where $H_i(\mathbf{b}) = 0$. And only Revised IP-OLDF is free from problem 1. If $\mathbf{x}_i$ is discriminated correctly, $e_i = 0$ and $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) >= 1$. If $\mathbf{x}_i$ is misclassified, $e_i = 1$ and $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) >= -9999$. It is expected that all misclassified cases will be extracted to alternative SV, such as $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) = -9999$. Therefore, the discriminant scores of misclassified cases become large and negative, and there are no cases where $y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) = 0$. Revised IP-OLDF can resolves first and second problems. Therefore, it is ready to be compared with other LFDs by 100-fold cross validation.

$$MIN = \Sigma e_i ; \quad y_i*(\mathbf{x}_i'\mathbf{b}+ b_0) >= 1 - M* e_i ; \quad (7)$$

b_0: free decision variables.

If e_i is a non-negative real variable, we utilize Revised LP-OLDF, which is an L1-norm LDF. Its elapsed runtime is faster than that of Revised IP-OLDF. If we choose a large positive number as the penalty c of S-SVM, the result is almost the same as that given by Revised LP-OLDF because the role of the first term of the objective value in equation (5) is ignored.

Revised IPLP-OLDF is a combined model of Revised LP-OLDF and Revised IP-OLDF. In the first step, Revised LP-OLDF is applied for all cases, and e_i is fixed to 0 for cases that are discriminated correctly by Revised LP-OLDF. In the second step, Revised IP-OLDF is applied for cases that are misclassified in the first step. Therefore, Revised IPLP-OLDF can obtain an estimate of MNM faster than Revised IP-OLDF [20].

It is regretful that all LDFs except for Revised IP-OLDF are not free from problem 1.

III. THE ROLL OF DATA IN THE RESEARCH

This basic research started after 1997 and ended in 2012. There are the following reasons why it needed sixteen years.

1) IP solver requested huge computation time before 2000 [20]. Therefore, it was too earlier to start from 1997.

2) IP-OLDF may not find true MNM if data is not in general position. This is not confirmed without the survey using the student data. Ibaraki and Muroga defined the same Revised IP-OLDF already [4]. But, it is very difficult to find mechanism why it can find the true MNM without the examination by real data and previous research of IP-OLDF.

In the first stage of this basic sturdy, the iris and CPD data were used for the evaluation of IP-OLDF and comparison with Fisher's LDF and QDF. IP-OLDF finds new facts such as: 1) the relation of NMs and LDFs, and 2) OCP, 3) MNM decreases monotonously. In the second stage, IP-OLDF finds that Swiss bank note data is linear separable. The student data reveals the defect of IP-OLDF that relates to problem 1. Even now, many researchers are not aware of this problem. Revised IP-OLDF resolved to find the interior point of the OCP directly.

After 2009, we started the applied research of linear separable data. I negotiated with the National Center for Univ. Entrance Examination (NCUEE), and got research data consisting of 105 exams in 14 subjects over three years. It was confirmed that error rates of LDFs except for Revised IP-OLDF cannot definitely recognize the linear separable data. More specifically, those of Fisher's LDF and QDF are very high. Eighteen pass/fail determinations of my statistical lectures are used for the research data. Tests have 100 items with 10 choice that are categorized four testlets scores such as: T1, T2, T3 and T4. If the pass mark is 50 points, a trivial LDF such as f = T1 + T2 + T3 + T4 - 50 can discriminate the pass/fail classes completely by the rule: $f \geq 0$ or $f < 0$. Students on the discriminant hyper-plane (f=0) are classified in the pass class because the discriminant rule is defined by four features definitely. Discrimination by 100 items finds serious problem 3 about the algorithm of the generalized inverse matrices of QDF. By the discrimination using four testlets, the error rates of Fisher's LDF and QDF are very high, and this is confirmed by 100-fold cross validation [15] [17].

IV. K-FOLD CROSS-VALIDATION

Re-sampling samples are generated from 4 real data sets. These are analyzed by 100-fold cross validation. Fisher's LDF and logistic regression are analyzed by JMP [7]. JMP division of SAS Institute Japan supports us to develop the program. Other LDFs are analyzed by LINGO [9]. LINDO Systems Inc. supports us to develop the program that is showed in [18][21]. The most important interest is the mean error rates of seven LDFs in the training and validation samples.

A. 100-fold cross validation of CPD

CPD data consisted of two classes: 180 pregnant women whose babies were born by natural delivery and 60 pregnant women whose babies were born by Caesarean section. There

 ISBN: 978-1-61208-358-2

were 19 features such as: X1 was the pregnant woman's age, X7 was the shortest anteroposterior distance, X8 was the fetal biparietal diameter, and X9 was X7-X8, X12 was X13-X14 (small normal random noise are added to X9 and X12), X13 was the area at the pelvic inlet, X14 was the area of the fetal head, and X19 was the lateral conjugate. X9 and X12 cause the multicolinearity. About 19 models selected by the forward stepwise method from 1-vat to 19-vars, NM of QDF is as follows: 22→20→22→18→18→16→15→9→9→ 8→9→ 21→17→16 →17→21→19→17→16. From 11–features to 12-features, NM increases 9 to 21 because X14 enter the 11-vars model and 12-vars model includes (X12, X13, X14). On the other hand, MNM decreases monotonously.

TABLE I. CPD DATA

OLDF	M1	M2
1-19	0.04	3.70
1-5,7-19	0.06	3.68
1,2,4,5,7-19	0.08	3.72
1,2,4,5,7-9,11-19	0.13	3.86
1,2,4,5,7,9,11-19	0.18	3.73
1,2,4,5,7,9,11-15,17-19	0.26	3.59
1,2,4,5,7,9,12-15,17-19	0.44	3.76
1,2,5,7,9,12-15,17-19	0.56	3.88
1,2,5,7,9,12,13,15,17-19	0.57	3.74
1,2,5,7,9,12,15,17-19	0.63	3.71
1,2,5,7,9,12,15,17-18	0.82	3.70
1,2,7,9,12,15,17-18	1.66	4.81
1,2,9,12,15,17-18	1.88	4.67
2,9,12,15,17-18	2.24	4.68
9,12,15,17-18	3.24	6.12
9,12,15,18	3.54	5.56
9,12,18	4.35	6.06
9,12	4.81	5.99
12	7.80	9.15
	M1Diff.	M2Diff.
SVM4	[0.08, 2.56]	[0.08, 1.07]
	0.45	0.39
SVM1	[1.03, 2.56]	[0.28, 1.43]
	1.76	1.43
LP	[0.07, 2.56]	[0, 1.28]
	0.45	0.28
IPLP	[-0.01(1), 0.05]	[-0.25(10), 0.11]
	0	0.02
Logistic	[0.18, 2.94]	[0.23, 1.52]
	0.97	0.63
LDF	[2.95, 7.69]	[1.68, 5.92]
	7.52	5.69

We examine 19 different models of CPD data selected by the forward stepwise method because there are over 500,000 models ($=2^{19}$-1). Table I shows the results by 100-fold cross validation. 'OLDF' is the result of Revised IP-OLDF. First column of 'OLDF' shows 19 models from 19-vars model to 1-var model. "M1 and M2" are the mean error rates for the training and validation samples. Those are computed by mean of 100 error rates of 19 different models. Therefore, M1 decreases monotonously as same as MNM. M1 of the full model is always minimum. The minimum value of M2 is 3.59% of 14-vars model. Therefore, we compare Revised IP-OLDF with six LDFs by this model.

"SVM4, SVM1, LP, IPLP, Logistic and LDF" are the results of S-SVM ($c=10^4$ and 1), Revised LP-OLDF, Revised IPLP-OLDF, logistic regression and Fisher's LDF, respectively. "M1 Diff. and M2Diff." are the difference of ((M1/M2 of six LDFs) - (M1/M2 of OLDF)). First row shows the ranges of 19 models. Second row shows the "M1Diff. & M2Diff." of the 14-vars model. "M2Diff." of LDF is 5.96%, and it is too bad. "M2Diff." of SVM1 is 1.43%, and it is worse than those of SVM4, LP and logistic. If we choose large value of penalty c such as 10000, the role of $\|\mathbf{b}\|^2/2$ in (5) is less meaning, and it may be similar to Revised LP-OLDF.

Only one 'M1Diff.' of IPLP is -0.01%. This means that Revised IPLP-OLDF is not free from problem 1 because M1 of Revised IP-OLDF is the minimum M1 among all LDFs. But ten 'M2Diff.' of IPLP are less than zero. Although some results may be caused by problem 1, other results may show that some M2 of Revised IPLP-OLDF are less than those of Revised IP-OLDF.

B. 100-fold cross validation of Iris data

Iris data [1] consisted of 100 cases with 4-features. Table II shows the results of 15 models by 100-fold cross validation. First column of OLDF shows the all possible combinations of features from 4-vars model (X1, X2, X3, X4) to 1-var model (X1). M1 of the full model is always minimum because M1 of (q+1)-features is always less than equal M1 of q-features theoretically. Although M2 of full model is minimum and it is 2.55, this is no guarantee theoretical. We consider the model with minimum M2 of Revised IP-OLDF as best model. Therefore, we can compere seven LDFs on this full model.

If we focus on 'M2Diff.' of the full model, those of SVM4, SVM1, LP, IPLP, Logistic and LDF are 0.46, 0.46, 0.4, 0.17, 0.39 and 0.64% worse than Revised IP-OLDF in the second row of SVM4, SVM1, LP, IPLP, Logistic and LDF. Results of six LDFs are not so bad because this data is very famous evaluation data of Fisher's LDF that satisfy the Fisher's assumption. Fisher chosen the best data for the validation of Fisher's LDF. Six maximum values of 'M2Diff.' are almost better than those 'M1Diff.'. This may imply that Revised IP-OLDF over-fit the training sample and the mean of error rates in the validation samples are worse than the training samples.

One 'M1Diff.' of LP and two 'M1Diff.' of IPLP are minus. This means that Revised LP-OLDF and Revised IPLP-OLDF are not free from problem 1 because M1 of Revised IP-OLDF is the minimum value among all LDFs. But, several M2 of Revised IP-OLDF are worse than others. Although some

 ISBN: 978-1-61208-358-2

results depend on the unresolved problem, other results may be caused by overestimate of Revise IP-OLDF.

TABLE II. IRIS DATA

OLDF	M1	M2
1,2,3,4	0.46	2.55
2,3,4	0.82	2.96
1,3,4	1.30	3.51
1,2,4	2.49	5.12
1,2,3	1.57	3.63
3,4	2.46	4.42
2,4	3.58	5.73
1,4	4.18	5.59
2,3	4.42	6.97
1,3	2.86	4.78
1,2	22.76	27.41
4	5.39	6.16
3	6.03	7.29
2	36.03	39.01
1	25.85	28.34
	M1Diff.	M2Diff.
SVM4	[0.58, 4.85]	[-0.67(3), 1.52]
	0.58	0.46
SVM1	[0.58, 4.85]	[-0.67(3), 1.52]
	0.58	0.46
LP	[-0.38(1), 3.63]	[-1.49(4), 1.43]
	0.47	0.40
IPLP	[-0.02(2), 0.07]	[-0.07(9), 0.17]
	0.01	0.17
Logistic	[0.71. 4.8]	[-1.11(3), 1.58]
	0.71	0.39
LDF	[0.51, 4.78]	[-0.89(3), 2.49]
	2.03	0.64

C. 100-fold cross validation of Swiss Bank data

Swiss bank note data consisted of two kinds of bills: 100 genuine and 100 counterfeit bills. There were six features. A total of 63 (=2^6-1) models were investigated. We had better considered about two types of discriminations: 16 linearly separable models and other 47 models. Table Ⅲ shows only 16 linear separable models in the training sample.

Only 4 M2s of Revised IP-OLDF are zero. In this case, we had better chosen the minimum number of features such as (1, 2, 4, 6). We compare Revised IP-OLDF with six LDFs in this 4-vars model. All models of six LDFs have the minimum M2. Those values are 0.38, 0.52, 0.27, 0.41, 0.38 and 0.47%, respectively. The results of six LDFs are not so bad. But, all M1s of SVM1 and Fisher's LDF are not zero. This means that both LDFs cannot recognize linear separable data, nevertheless this data may satisfy Fisher's assumption because genuine/counterfeit bills are industry products. And the results of H-SVM are as same as SVM4.

TABLE III. SWISS BANK NOTE DATA

OLDF	M1	M2
1,2,3,4,5,6	0.00	0.00
2,3,4,5,6	0.00	0.24
1,2,4,5,6	0.00	0.00
1,2,3,4,6	0.00	0.00
1,2,3,5,6	0.00	0.10
2,4,5,6	0.00	0.21
2,3,4,6	0.00	0.16
1,2,4,6	0.00	0.00
2,3,5,6	0.00	0.03
1,2,3,6	0.00	0.08
1,2,5,6	0.00	0.10
2,4,6	0.00	0.15
2,3,6	0.00	0.02
2.5,6	0.00	0.02
1,2,6	0.00	0.03
2,6	0.00	0.01
	M1Diff.	M2Diff.
SVM4	0	[0.22, 0.68]
/ HSVM	0	0.38
SVM1	[0.24, 0.57]	[1, 2.5]
	0.27	0.52
LP	0	[0.27, 0.75]
	0	0.27
IPLP	0	[0.25, 0.75]
	0	0.41
Logistic	0	[0.22, 0.65]
	0	0.38
LDF	[0.44, 0.95]	[0.25, 1.02]
	0.44	0.47

D. 100-fold cross validation of Student data

The student data consists of two groups: 25 students who pass the exam and 15 students who fail. There were 3 features: X1 was the hours of study per day; X2 was spending money per month; X3 was number of days drinking per week. If we analyze 2-features (X1, X2) by IP-OLDF, IP-OLDF chosen X2=5 as the discriminant hyper-plane. Four pass students and four fail students spent 50,000 yen/month. These eight students are on the discriminant hyper-plane. Only three fail students who spent less than 50,000 yen are misclassified by IP-OLDF. Revised IP-OLDF finds three true MNMs are 5 using by LINGO k-best option [21].

We examine 7 different models of student data by 100-fold cross validation in the Table Ⅳ. M1 decreases monotonously from 1-var to 3-vars. There are 6 passes such as : 1→(1,2)/(1,3)→(1,2,3), 2→(1, 2)/(2, 3) →(1, 2, 3), 3→

 ISBN: 978-1-61208-358-2

(1,3)/(2, 3) →(1,2,3). We compare Revised IP-OLDF with six LDFs by the model (2, 3). 'M2Diff.' of LDF, logistic, SVM4, SVM1 and LP are 7.19, 5.88, 4.1, 4.1 and 3.23%. These are very bad. 'M2Diff.' of Revised IPLP-OLDF is -0.2%. This may be the defect of Revised IP-OLDF.

TABLE IV. STUDENT DATA

OLDF	M1	M2
1,2,3	5.70	12.78
1,2	9.18	15.15
1,3	10.30	15.45
2,3	7.45	9.05
1	16.43	19.10
2	14.68	17.63
3	17.75	21.23
	M1Diff.	M2Diff.
SVM4	[1.2, 4.15]	[-0.7(2), 4.1]
	3.2	4.1
SVM1	[1.2, 4.15]	[-0.7(2), 4.1]
	3.2	4.1
LP	[-0.29(3), 3.35]	[-4.53(3), 3.23]
	2.7	3.23
IPLP	[0, 0.2]	[-0.75(5), .25]
0	0	-0.2
Logistic	[0.83, 5.43]	[-1.45(3), 5.88]
	4.8	5.88
LDF	[2.55, 6.55]	[-1.15(1), 7.19]
	6.03	7.19

V. CONCLUSION

Many statisticians believe that MNM criterion is foolish criterion because it over-fit for the training sample and it may overestimate the validation sample. On the contrary, generalization ability of LDF is best because it follows the normal distribution without examination by real data. In our paper, this claim may be wrong. In near future, this will be confirmed by the discrimination of the linear separable data using the pass/fail determination. In addition, the mean error rates of Fisher's LDF are higher than other LDFs. Past important researches using LDF should be reviewed, especially in the medical diagnosis. K-fold cross validation is very useful, compared with the leave-one-out method [5].

ACKNOWLEDGMENT

My research was achieved by LINGO of LINDO Systems Inc., and JMP of SAS Institute Inc.

REFERENCES

[1] A. Edgar, "The irises of the Gaspe Peninsula," Bulltin of the American Iris Society, vol. 59, pp. 2-5, 1945.

[2] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 7, pp. 179–188, 1936.

[3] B. Flury and H. Rieduyl, Multivariate Statistics: A Practical Approach. Cambridge University Press, 1988.

[4] T. Ibaraki and S. Muroga, "Adaptive linear classifier by linear programming," IEEE transaction On systems science and cybernetics, SSC-6, pp. 53-62, 1970.

[5] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," Technometrics vol. 10, pp.1-11, 1968.

[6] J. P. Sall, SAS Regression Applications. SAS Institute Inc. 1981.

[7] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics, 3rd ed. SAS Institute Inc. 2004.

[8] L. Schrage, LINDO－An Optimization Modeling System－. The Scientific Press. 1991.

[9] L. Schrage, Optimization Modeling with LINGO. LINDO Systems Inc. 2006.

[10] S. Shinmura and A. Miyake, "Optimal linear discriminant functions and their application," COMPSAC79, pp. 167-172, 1979.

[11] A. Miyake and S. Shinmura, "An Algorithm for the Optimal Linear Discriminant Function and its Application," Japanese Journal of Medical Electronics and Biological Engineering, Vol 18/1, pp. 15-20, Feb. 1980.

[12] S. Shinmura, "Optimal Linear Discrimrnant Functions using Mathematical Programming," Journal of the Japanese Society of Computer Statistics, vol. 11/2, pp. 89-101, 1998.

[13] S. Shinmura, "A new algorithm of the linear discriminant function using integer programming," New Trends in Probability and Statistics, vol. 5, pp. 133-142. 2000.

[14] S. Shinmura, The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing. 2010.

[15] S. Shinmura, "Beyond Fisher's Linear Discriminant Analysis - New World of Discriminant Analysis -," 2011 ISI CD-ROM, pp.1-6. 2011.

[16] S. Shinmura, "Problems of Discriminant Analysis by Mark Sense Test Data," Japanese Society of Applied Statistics, vol. 40/3, pp. 157-172, 2011.

[17] S. Shinmura, "Evaluation of Optimal Linear Discriminant Function by 100-fold cross-validation," 2013 ISI CD-ROM, pp.1-6, 2013.

[18] S. Shinmura, "Evaluation of Revised IP-OLDF with S-SVM, LDF and logistic regression by K-fold cross-validation," IEICE Technical Report IBISML 2013-44 (2013-11), pp.61-68.

[19] S. Shinmura, "End of Discriminant Functions based on Variance-Covariance Matrices," ICORE2014, pp. 5-16. 2014.

[20] S. Shinmura, "Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP," Statistics, Optimization and Information Computing, vol. 2, June 2014, pp 114-129.

[21] S. Shinmura, "Three Serious Problems and New Facts of the Discriminant Analysis" Operations Research snd Enterprise Systems, ICORES 2014, Revised Selected Papers, in Press.

[22] A. Stam, "Nontraditinal approaches to statistical classification: Some perspectives on Lp-norm methods," Annals of Operations Research, vol. 74, pp. 1-36, 1997.

[23] V. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag. 1995.

[24] R. Warmack and R. C. Gonzalez, "An Algorithm for the optimal solution of linear inequalities and its application to pattern recognition," IEEE Trans. Computers, pp. 1065-1075, 1973.

 ISBN: 978-1-61208-358-2

Use of Social Microblogging to Motivate Young People (NEETs) to Participate In Distance Education Through www.eBig3.eu

Dace Ratniece
Liepaja University, Latvia
ratniece.dace@gmail.com

Abstract -Young people, who are the fundamental asset of our economies and societies across the world, face a real and increasing difficulty in finding a decent job with each day. Three additional merging factors are worsening the youth employment crisis even further, causing challenges while transiting to decent jobs, namely (i) numbers of discouraged youth, in other words, young people, who are neither in education nor in employment or training (NEETs) are increasing, (ii) unemployment among university graduates of tertiary education in general are rising and (iii) potential NEET group students, especially in the 1st year, who, apart from reduced study fees, require extra motivation and moral support from educators. What can be done to solve the dropout crisis with microblogging possibilities? Many educators have found that microblog Twitter is a great help tool to increase student participation and further engagement in lifelong education once they have left the classroom. The use of e-learning in itself does not constitute an enhancement of the quality of teaching and learning, but it is a potential enabler for such enhancement. The provided examples of good practice illustrate how certain initiatives address challenges and seek to effectively engage learners. The study aim - to make young people NEET problem analysis in order to understand their motivation to engage in distance learning process. The first insurance against unemployment is quality education and training. Distance learning project eBig3 is example of unique way to promote reintegrating young people into the labour market and education too.

Keywords–youth; NEET group; microblogging; Twitter; e-learning; distance education; www.eBig3.eu.

I. INTRODUCTION

The immediate future of Europe depends upon 94 million of Europeans aged between 15 and 29. Apart from the challenges that young people have been facing for generations while embarking upon adult life, this generation lives in the era of all embracing globalization and has to cope with responsibility for ageing population. So it is a matter of great concern that these young people have been hit so hard by the economic crisis, said at the document "*The Youth Employment crisis: Highlights of the 2012 ILC Report, International Labour Office (ILO), 2012"* [1]. In 2011, only 34% of young people were employed; this is the lowest figure that Eurostat has ever recorded. Unemployment figures also testify significant difficulties that young people face while entering the labour market; since the start of recession, youth unemployment has increased by 1.5 million reaching 5.5 million (21%) in 2011.

Serious as these statistics may be, they do not adequately capture the situation of young people because many are students, and hence, are classified as out of the labour force. For this reason, European Union (EU) policymakers have been increasingly using the NEETs concept: 'not in employment, education or training'. This concept is included in document "*NEETs: Young people not in employment, education or training: Characteristics, costs and policy responses in Europe, 2012 "[2]*. In principle, the definition is straightforward; it refers to those who currently do not have a job, are not enrolled in training or are not classified as students. It is the measure of disengagement from the labour market and perhaps from society in general.

The European Parliament has expressed 'serious doubts' about the scale of the actions proposed by the European Commission (EC) to address a high rate of youth unemployment that the EU is facing now when the average rate of youth unemployment has reached 23.7% and is affecting many member states. Since 2010, on the initiative of the Committee on Employment and Social Affairs and the Committee on Culture and Education of the Parliament, practical actions aimed at promoting youth employment have been regularly proposed. All recommendations of the Parliament will be included in the report *Youth Unemployment: Possible Ways Out* prepared by Joanna Katarzyna Skrzydlewska (2013/2045 (INI) (in progress).
One of the main targets of the Europe 2020 Strategy is that the share of early school leavers should be under 10%.

Unfortunately, youth policy is still not high priority in Latvia, although the EC has already adopted a very important document, the European Parliament resolution of 18 May 2010 on *"An EU Strategy for Youth – Investing and Empowering"* (2009/2159(INI).

 ISBN: 978-1-61208-358-2

This report on the Baltic youth policy in line with the EU strategy for youth, as a contribution to the creation of the Baltic youth policy statement, was published.

Report-research (100 pages and references) on youth policy in the Baltic states was drafted by Dace Ratniece [3], in 2012.

According to ILO data for 2011, there are 1.2 billion young people between the ages of 15 and 24 in the world and about 90% of them live in developing countries. The current population of young people is the largest the world has ever seen [1].

Let us look at some data on youth unemployment in Latvia.

January 2013: The total number of unemployed people in Latvia was 107.488. 11.001 (10%) of all unemployed people are young people between the ages of 15 and 24. The majority of them, 7.074 (64%), have been unemployed for up to six months.

Long-term unemployment in Latvia is continuing to rise. Do we have any specific ideas how to save this "lost generation", while there is still time to do so?

A. The aim of the study

The aim of this study is to identify potential scenario of distance learning and self-development of youth using a variety of collaborative and motivational approaches, microblogging and social networking on Twitter and the digital environment as distance learning tools.

B. Research object

The object of this research is to explore motivation that enhances advancement of young people in the NEETs group and participation in the distance learning process. Social network microblog Twitter could become a successful digital media tool to involve NEETs in distance learning by using appropriate pedagogical and psychological methods of motivation.

C. Objectives of the research

- To analyze philosophical, pedagogical and psychological literature as well as articles on computer science that evaluate e-learning.
- To study self-expression and self-development of young people microblogging on Twitter.
- To define pedagogical and psychological techniques of motivating Twitter users NEETs to participate in distance learning.
- To identify drawbacks of distance learning courses.
- To develop distance learning courses by taking into account the needs of NEETs and thus ensure their integration into society.

D. Expected results

- Self-expression skills of NEETs will increase and that will lead young people to further professional development (microblogging on Twitter and digital distance learning tools).
- Self-development capabilities of NEETs will increase.
- NEETs will become more aware of their own potential, belief in their own abilities will increase and their motivation to study and develop further will be stimulated.

After this introductory part, we present in Section II the general position of our work in the field of NEETs problems solving through cyberspace. In Section III we present results. Section IV concludes the paper.

II. GENERAL POSITION

While preparing the report [3], we have drawn particular attention to the issue of young people in the NEET group. Young people are giving up on the job search altogether because of low prospects of ever finding a job during the crisis or low prospects of finding a decent job according to their skills. School dropouts exposed to gang culture and drugs at a young age or young people from deprived socio-economic backgrounds also face diminished chances of gaining employment.

Unemployment among university graduates, graduates of tertiary education institutions in general has been rising [1]. That is either because of the deterioration of education standards or a mismatch between graduates' skills and available jobs. This phenomenon causes concern for several reasons:

- Contradicts the assumption that higher education increases employability.
- Suggests that high-cost investment into higher education is wasteful because social return from unemployed graduates is low.
- Increases "brain drain" of skilled youth from many developing economies.
- Causes personal and political frustration.

Some important features that NEETs share:

1) They have gained skills and knowledge outside the formal system.

2) They more often tend to be at disadvantage in terms of, for example, capabilities, level of education, background (imigrant families) or economic conditions.

3) They more often tend to remain unemployed, their integration into the labor market is more difficult, concludes Researcher G.Pollock at the article "Youth Transitions: Debates over the social context of Becoming an adult [4].

Cyberspace helps young people find someone to share their interests and needs with. It is the new medium of

 ISBN: 978-1-61208-358-2

communication and parents are often unaware of what information teens share on social media. Teenagers explore their identities, experiment with behavioral norms, date and build friendships Correa at al. [5].

Many young people like microblogging on Twitter. Twitter it is one of the fastest growing social networks in the world. What is better than to post information in 140-character tweets? Compared to other complicated platforms, Twitter is engaging and simple. Its development was inspired by cell phone text messaging and its creators instinctively or knowingly became the inventors of the most uncomplicated social media systems said in the project *"Our Featured Projects, NJI Media"*[6].

Twitter provides a social platform for discovery of new connections, weak links according to sociologist Mark Granovetter. If Granovetter is right, it is through weak links that we learn about new ideas, are exposed to different concepts and our status quo is shaken up. Twitter is characterized as a simple tool for personal or professional use, an outlet for keeping up with the news, sharing of information, crowdsourcing [7].

D. Ratniece started microblogging on Twitter and got involved in problems that NEETs face. This inspiration came from analyzing the aspects of microblogging Y. Amichai-Hamburger [8]. For good microblogging one firstly needs to know how to build a profile. The ideas of profile building are as follows [9]:

1) *Create a unique identity*: a Twitter profile is very important. How to make tweets more personable and human? People are more likely to communicate with real people so you should add a little of your character to your tweets and make some genuine connections. Establishing a personality means creating a character, if you have not got one yet. It is important to have a good profile page with a Picture and a descriptive bio *(Fig. 1.)*

2) Tweet regularly: tweeting on a regular basis will keep you in contact with your followers and that will also keep your page active. The optimum number of tweets per day may vary; but daily tweeting, 1-4 tweets per day, works for most users.

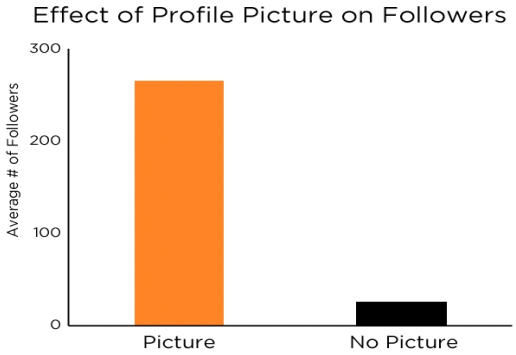


Figure 1. Effect of a profile picture on followers [10]

3) *Encourage your followers to retweet:* it is not always easy to get your information retweeted by other users but there are a few ways how to maximize. Use a Twitter Timeline to analyze your most successful retweeting periods, which could indicate the best time for you to tweet. Interacting with other users will help you build beneficial relationships and those who consider you a friend are probably more likely to retweet your information on a regular basis.

4) *Be interesting:* if you want to have more followers who care about you and what you share, make sure your tweets are interesting. Use «Tweeteffect» to get a detailed look at which of your tweets are gaining followers and which are causing to lose them. The result could be random or you could discover a pattern.

5) Interact and share: if you want people to share your tweets, you have to return the favor. Interact and share with other people and you will get 'thanks' in return. Join in big conversations and use the hash tag (#) to take part in trending topics but always stay relevant to maintain follower retention.

Given the above conditions, D. Ratniece [11] started talking to young people, microblogging on Twitter, and invited them to start training in distance education. Microblogging on Twitter attracts young digital media communicators and that may be used to identify and motivate youngsters, youth in the NEET group aged 13-25 years, to participate in distance education. Understanding the problems of NEETs may lead to solutions for distance learning and it would be more likely to generate involvement when free distance learning courses are offered by Kapenieks [12] and Ratniece [11].

EC report „Opening up education: innovative teaching and learning for everyone through the use of new technologies and open educational resources" [13] highlights that nowadays students need more personalized contact with teachers, greater cooperation and better links between formal and non-formal education, which can largely be implemented by school learning using digital technologies. In this respect the EU risks falling further behind other regions of the world.

In order to transform their education and training, the United States and some Asian countries are investing in ICT strategies. These countries are transforming their education systems by modernizing them and making them accessible internationally, thus achieving remarkable results regarding the accessibility to education and study fees, pedagogical practice and gaining of a global recognition. For example, a large proportion of digital content is provided by the market participants from outside the Europe, including educational institutions that offer training programs on a global scale through massive open online courses (MOOC) [12].

There is a particular need to worry about third group – potential NEET group students, especially in the 1st year,

 ISBN: 978-1-61208-358-2

who, apart from reduced study fees, also require extra motivation and moral support from educators.

Therefore, it is necessary to deal with the three groups of young people NEET:

1) Those who belong to NEET group.
2) Those who discontinued their university studies due to financial reasons and started a full-time job.
3) Potential NEET group of universities students, especially in the 1st year.

All three target groups share the same need: to gain experience or receive educational opportunities in order to get a steady job and sustain themselves financially. Are there any initiatives to address this situation, and if so, have they been successful? Of course, there are also extreme measures, such as migration, but would it help in the long run?

That is the case of the Latvian-Lithuanian cross-border project *eBig3* (Fig. 2). The project aims to create a network for cross-border research cooperation in technology enhanced learning (TEL) and to develop a strategy for educational business promotion service. The Project combines three kinds of TEL in a complementary way:

1) E-learning – mainly computer and/or internet-based learning.

2) T-learning: TV based learning.

3) M-learning: learning with a use of mobile devices; to produce an effective and innovative cross-media learning delivery system (eBig3) that goes beyond traditional web-based learning approaches.

This is an innovative project of open and distance learning. The developed solutions includes integration of technical issues for cross-media learning content delivery, refinement of pedagogic considerations, development of shared understanding of target users learning contexts in border areas, production of learning content and organizing course pilots. This is a unique way to promote reintegrating young people into the labour market and education too.

Figure 2. Latvian-Lithuanian cross-border project eBig3

III. RESULTS

We participated in 2013 in the examination of 1st year students of program “Telecommunications (Distance Learning e-course) of Riga Technical University (RTU).

During the exam, the author conducted a survey, asking 107 students to fill in a questionnaire - Assessment of the effectiveness of the learning methods practiced on a scale from 1 (the lowest) to 10 (the highest), and a short commentary. Table 1 of this study shows the students’ evaluation of the efficiency of the learning form used in this course which reflects students’ support for both traditional forms - learning and e-learning.

Respondents very carefully evaluated both the negative and the positive aspects of both methods based on their personal experience, and were able to provide an objective feedback in regards to what situations required a direct contact with teacher, and when e-learning was the best and most efficient learning option. That is reflected in the questionnaires that were submitted at the end of semester:

"E-learning is not a substitute for the traditional learning methods at the moment, but it can be a good additional tool for learning."

"E-learning is very useful when one needs to obtain lecture files and other study materials, and it’s especially great for those who cannot attend lectures because of work or other commitments."

"I find E-learning very convenient because of its reduced costs and innovative learning environment. Nevertheless, it is also important to keep the traditional aspects of learning as a part of it in order to maintain the diversity of the course”.

"Traditional forms of studies are definitely very important, because the presence of a teacher can contribute to a better understanding of the subject."

"E-learning should be available for all the subjects, because it is very convenient."

Respondents indicate that e-learning and traditional forms of study need to be kept in balance, because e-learning provides a great advantage to learn anywhere, anytime. A successful guidance through the study process, however, is just as important, and can only be ensured when a teacher is present. Respondents were also asked to state the core competences of the teacher:

1) responsive, intelligent, able to establish a good contact with students; sociable, friendly yet demanding when it comes to the quality of students’ performance; understanding, able to listen and motivate their students;

2) possesses the ability to initiate discussions, to prepare a training plan, has a comprehensive understanding of different study subjects, ability to keep students motivated and interested in the particular subject of study;

 ISBN: 978-1-61208-358-2

TABLE I STUDENTS EVALUATION OF THE EFFECTIVENESS OF THE FORM

	Low rating			Average rating				High rating		
Form of study	1	2	3	4	5	6	7	8	9	10
Lectures				3	6	16	34	29	11	6
Discussions			2	3	5	13	18	35	21	8
Assignment preparation and insertion in ORTUS system*			1		7	11	19	35	15	17
Teachers' comments in ORTUS system			1	2	7	12	15	33	17	18
e-portfolio utility	9	5	6	3	10	16	26	16	10	4

* RTU portal «ORTUS» (www.ortus.lv), which provides e-learning environment

According to the students, just as important as experience in their field and professionalism, are positive traits of character, suggesting the need to develop a positive communication between the students and the teacher.

IV. CONCLUSIONS AND FUTURE WORK

1. Being NEET is not only a personal problem for those affected, but constitutes a challenge to society as a whole. This is very important given the size of the NEET population today, which may seriously undermine the sustainability and stability of the societies concerned. Governments have been very active in promoting policies for re-engaging young people in the labour market and the education system.

2. In addition to increasing access to education, greater use of new technologies and open educational resources can also help reduce the costs of educational institutions and students, particularly for disadvantaged groups, including existing and potential NEET group of people.

3. Educational institutions should combine the traditional forms of study and e-learning, as e-learning provides the opportunity to learn anywhere, anytime. However, a direct contact between the teacher and the student plays a very important role in acquiring a quality education, and should by no means be left out.

4. Latvian-Lithuanian cross-border project eBig3 is a unique way to promote reintegration of young people into the labour market and education. It is therefore necessary to develop distance learning courses by taking into account the needs of NEETs and thus ensure their integration into society. Social network microblog Twitter could become a successful digital media tool to involve NEETs in distance learning by using appropriate pedagogical and psychological methods of motivation.

REFERENCES

[1] The Youth Employment Crisis: Highlights of the 2012 ILC Report, Geneva, International Labour Office, 2012, pp.44 https://www.google.lv/#q=The+Youth+Employment+Crisis%3A+Highlights+of+the+2012+ILC+Report [retrieved: 08.2014]

[2] NEETs: Young people not in employment, education or training: Characteristics, costs and policy responses in Europe, Luxembourg: Publications Office of the European Union, 2012, pp.171, file:///C:/Users/Admin/Downloads/EF1254EN.pdf http://www.eurofound.europa.eu/pubdocs/2012/54/en/1/EF1254EN.pdf [retrieved: 08.2014)

[3] D. Ratniece, "Report-research on the Baltic youth policy in line with the EU strategy for youth as a contribution to the creation of the Baltic youth policy statement", 2012, pp.100 (unpublished).

[4] G. Pollock, "Youth transitions: Debates over the social context of becoming an adult", Sociology Compass, 2008, 2(2), pp. 467–484.

[5] T. Correa, W.H. Amber, and H.G. de Zúñiga, "Who interacts on theWeb?: The intersection of users' personality and social media use", Center for Journalism & Communication Research, School of Journalism,University of Texas at Austin, USA, Computers in Human Behavior, 2010, 26, pp. 247–253.

[6] Our Featured Projects, NJI Media.. http://njimedia.com/infographic-5-12-best-twitter-practices/ [retrieved: 08.2014]

[7] M.S. Granovetter, "The Strength of Week Ties", American Journal of Sociology, Vol. 78, Issue 6, May, 1973, pp. 1360-1380. http://sociology.stanford.edu/people/mgranovetter/documents/granstrengthweakties.pdf. [retrieved: 08.2014]

[8] Y. Amichai-Hamburger, "Internet and personality", Computers in Human Behavior, 2002, 18(1), pp. 1–10.

[9] N.Arceneaux, A.Schmitz Weiss, "Seems stupid until you try it: press coverage of Twitter", 2006–9, New Media & Society, 2010 12(8), pp. 1262–1279.

[10] D. Zarella, Twitter accounts with a profile picture have 10 times more followers than those without, 2010 http://blog.hubspot.com/blog/tabid/6307/bid/5811/Twitter-Accounts-with-a-Profile-Picture-Have-10-Times-More-Followers-Than-Those-Without.aspx [retrieved: 08.2014]

[11] D. Ratniece, "Social microblog TWITTER use to motivate young people (NEETs) to involve in distance education". Proc. of the international Scientific Conference "Society, Integration, Education," May 24th – 25th, 2013 (Rezekne Higher Education Institution, Faculty of Education and Design, Personality Socialization Research Institute in Collaboration with University of Udine, Italy), Vol. II, Rezekne: Rezekne Higher Education Institution, 2013, pp.449-464, ISSN 1691-5887. [http://www.ru.lv/ckfinder/userfiles/RAweb/Saturs/zinatne/zinatniskie_instituti/personas_socializacijas_petijumu_instituts/izdevumi/2013/II%20da%C4%BCa.pdf [retrieved: 08.2014]

[12] A. Kapenieks, B. Žuga, G. Štāle and M. Jirgensons, "E-ecosystem Driven E-learning vs Technology Driven E-learning", Proc. of 4th International Conference on Computer Supported Education (CSEDU 2012). Portugal, Porto, 16-18 April, 2012, pp. 436-439.

[13] Opening up education: innovative teaching and learning for everyone through the use of new technologies and open educational resources"/COM/2013/0654, Brussels. pp. 13 http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013DC0654&from=EN [retrieved 08.2014]

 ISBN: 978-1-61208-358-2

Circadian Patterns in Twitter

Marijn ten Thij, Sandjai Bhulai

VU University Amsterdam,
Faculty of Sciences,
Amsterdam, The Netherlands
Email: {m.c.ten.thij,s.bhulai}@vu.nl

Peter Kampstra

RTreporter,
Amsterdam, The Netherlands
Email: peter@rtreporter.com

***Abstract*—In this paper, we study activity on the microblogging platform Twitter. We analyse two separate aspects of activity on Twitter. First, we analyse the daily and weekly number of posts, through which we find clear circadian (daily) patterns emerging in the use of Twitter for multiple languages. We see that both the number of tweets and the daily and weekly activity patterns differ between languages. Second, we analyse the progression of individual tweets through retweets in the Twittersphere. We find that the size of these progressions follow a power-law distribution. Furthermore, we build an algorithm to analyse the actual structure of the progressions and use this algorithm on a limited set of tweets. We find that retweet trees show a star-like structure.**

Keywords–Data analytics; Twitter; Retweet graph; Language use; Daily pattern

I. INTRODUCTION

In the current digital age, many different (micro)blogging platforms have emerged, e.g., Twitter, Facebook and LinkedIn. Using these media, users can share everyday thoughts and activities. Through this sharing behaviour, large quantities of information are available to researchers who have distinguished many applications for the valorisation of this information. Within these applications, the focus is placed on predicting the future. The implementation of predictions using data from Twitter cover very different areas, e.g., predictions with respect to policital elections [1], the prediction of stock market prices [2], estimating the box-office revenue of a movie [3], and the detection of earthquakes [4].

In most social media, there is a notion of trending topics. With this notion, questions arise as to how and when these topics emerge and grow. Since these questions are far from trivial, we need more insight in the normal use of this social media to be able to answer these questions. Because if one understands the patterns of usage of a social medium, this insight will lead to understanding of the trending mechanism.

Therefore, in this paper, we study the patterns of use in Twitter. We focus on two aspects that provide a first insight in the usage of Twitter. First, we give the reader an overview of related research in Section II. Then, in Section III, we describe the process of gathering the data we used in the study. The analysis of circadian patterns in Twitter is presented in Section IV. Thereafter, in Section V, we display our analysis of the progression of retweets through the user network of Twitter. Finally, we draw our conclusions and discuss possible extensions of our work in Section VI.

II. RELATED RESEARCH

In this section, we provide the reader with a brief overview of related research in the two areas that we adress in this paper. The first of these fields is the circadian pattern that appears in social media usage. Secondly, we focus on the spread of information in social media.

A. Circadian patterns in Social Media

First, we consider social media activity. This has been researched for many different social media platforms. For instance, Kaltenbrunner et al. [5] analyse the activity pattern in Slashdot, a news site. They observe both daily and weekly activity patterns in the use of the site. Also, Gill et al. [6] study the activity of Youtube. They find clear circadian and weekly patterns, where the majority of the activity takes place at the end of the day during weekdays. Szabo and Huberman [7] examine the activity patterns on Digg and Youtube. They notice a weekly cycle of activity in Digg and investigate the popularity pattern of articles in Digg and videos in Youtube.

Noulas et al. [8] investigate the user activity pattern on Foursquare. They find clear geo-temporal rhythms in its activity, both for weekdays and weekends. Moreover, Grinberg et al. [9] use Foursquare data to extract real-life activity patterns. They observe that there are clear patterns for coarse categories, such as food or nightlife. They also notice that these patterns are present in Twitter.

Yasseri et al. [10] analyse circadian patterns in editorial activity on Wikipedia, an online encyclopedia. They find a clear daily pattern in activity per language. The only exception to these patterns is the English Wikipedia. For this language, the activity is more spread out over the day. Also, they find four weekly activity patterns for different groups of countries. Ten Thij et al. [11] study the page-view activity patterns for Wikipedia and observe circadian patterns in page-view activity.

Poblete et al. [12] investigate the user activity on Twitter for the top 10 countries in their sample. They perform an analysis of the activity based on sentiment and network properties. They find that the network and user properties can differ from country to country, from small connected networks to a large and more hierarchical structured network. Mocanu et al. [13] study GPS-tagged tweets by location and language. They analyse the heterogeneity of language use for many levels (e.g., global, country and city) and observe clear peaks in activity by tourists in some countries in the Mediterranian.

 ISBN: 978-1-61208-358-2

B. Information spread

In the second part of our work, we focus on the spread of information through the network. Again, this type of activity analysis has been executed for multiple platforms. For instance, Jurgens and Lu [14] analyse temporal patterns in edits to Wikipedia articles. Their analysis reveals motif instances in the edit-patterns to pages.

Lerman and Gosh [15] analyse user activity on Digg and Twitter. They conclude that despite their different setup, both sites display similar patterns of information spread. Kamath et al. [16] analyse the geo-spacial progression of hashtags in Twitter using geo-tagged tweets. They use their analysis to find analytics techniques to characterize the relative impact of locations on spread dynamics of a topic. Yang and Leskovec [17] investigate temporal patterns arising in the popularity of online content. They formulate this as a time series clustering problem and formulate an algorithm to cluster these time series with respect to the patterns they exhibit. Finally, Bhamidi et al. [18] develop a random graph model that models the giant component of the retweet network induced by an event on Twitter. We aim to extend this insight to a message-based insight in the spread dynamics of Twitter.

III. DATASET

In this section, we describe how we obtained the tweets that were using in our analysis. The tweets were scraped by RTreporter, a company that uses the incoming stream of Dutch tweets to detect news in the Netherlands. These tweets are scraped using the filter stream of the Twitter Application Programming Interface (API) [19]. We set up four different streams, the first two streams (called sample and geo-located) are meant to give an overview of the stream of Twitter messages. The third and fourth stream (called "Netherlands (NL) general" and "NL specific") are set up with the goal to scrape as many Dutch tweets as possible.

The first stream is the so-called sample stream. It outputs a sample of the complete Twitter Firehose. The sample that is given, contains roughly 1% of all tweets. The second stream uses the option **location** of the filter stream, in which a geo-location square is defined. All the tweets within this square are caught. We filter the stream on the geo-square induced by ((-179.99, -89.99), (179.99, 89.99)).

We call the third stream the "NL general" stream. In this stream, we use filter stream with the option **track**, where a list of words must be defined. All tweets containing one of these words are caught. We define a list of general Dutch words (e.g., *'een, het, ik, niet, maar, heb, jij, nog, bij'*). In total, this list consists of 130 words. Lastly, the fourth stream, "NL specific" uses the filter stream combining **track** and **follow**. For this last option, we add a list of user IDs for which all tweets are caught. In total, we define a list of 1,303 users. Examples of accounts are *@NUnl,@TilburginBeeld*. The list of terms consists of 395 entries (e.g., *'brandweer, politie, gewond, ambulance'*). Note that a specific tweet may be contained in multiple streams, we have not distinguished duplicates in our dataset.

Since we do not have access to the Twitter Firehose, we do not receive all tweets that we request due to rate limitations by Twitter [20]. An overview of the missed tweets is presented in Table I. Furthermore, in Figure 1, we display the number of tweets that were missed per stream on a daily scale. We see that the number of missed tweets in the 'NL general' stream is gradually decaying over time. In our experience, this decrease is probably caused by a decrease in the number of 'spam' tweets made by Dutch teens (e.g., tweets like "Welterusten!", which means "Good night!"). The geo-located stream follows an increasing trend. Furthermore, the sample stream has no missed tweets, which is logical, since the maximum number of tweets that one can receive is bounded by this number. With respect to the 'NL specific' stream, we see that the number of missed tweets is small, with the exception of some dates. We have not performed a more detailed analysis of the specific activity during these days.

TABLE I: NUMBER OF MISSED TWEETS PER STREAM FROM FEBRUARY 1ST 2013 TO FEBRUARY 1ST 2014.

'NL general'	Sample	Geo-located	'NL specific'
58,711,359	0	1,744,800,984	10,025

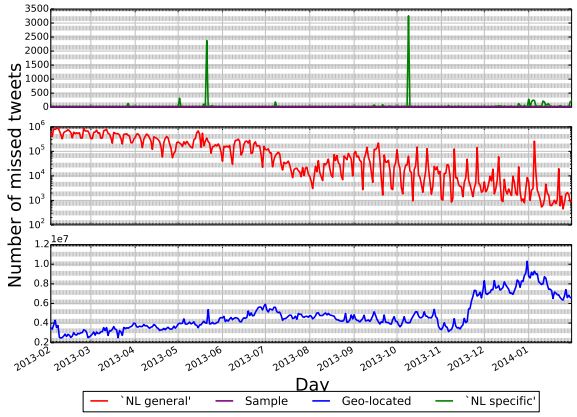


Figure 1: Number of missed tweets per day, indicated per stream.

We process the tweets that have been obtained by these four streams from February 1st 2013 to February 1st 2014. The number of tweets, clustered by language, for each stream is indicated in Table II. For the sample stream, we also indicate which percentage of all tweets was posted in the given language (denoted in brackets behind the number of tweets). The timezone information displayed in Table II is used to correct the daily/weekly patterns for comparison. If a language is commonly used in multiple timezones, we search for a city that is located in the center of this area and base the correction on this city. These cities are also mentioned in Table II.

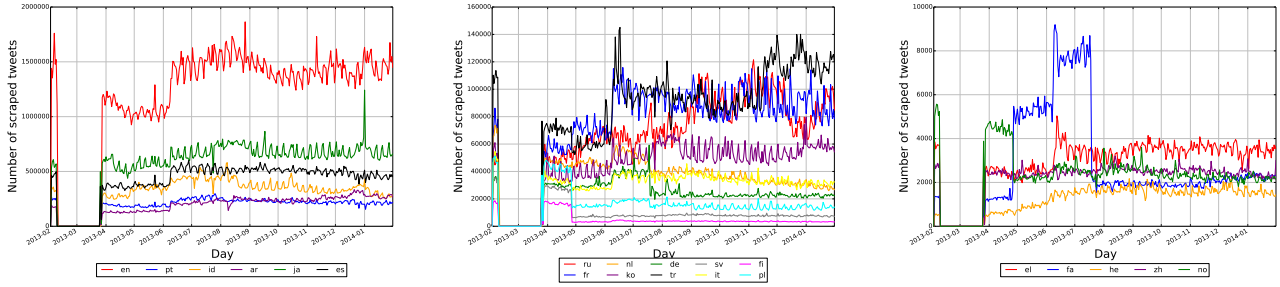
IV. TWEET PATTERNS

In this section, we discuss the temporal patterns that emerge in the data. We focus on two streams, namely the geo-located and sample streams, since these streams give a wide perspective on the traffic on Twitter. We present an analysis of these two streams on a daily scale for the complete year. Also, we present a more fine-grained analysis of the hourly patterns, both on a daily and a weekly basis.

 ISBN: 978-1-61208-358-2

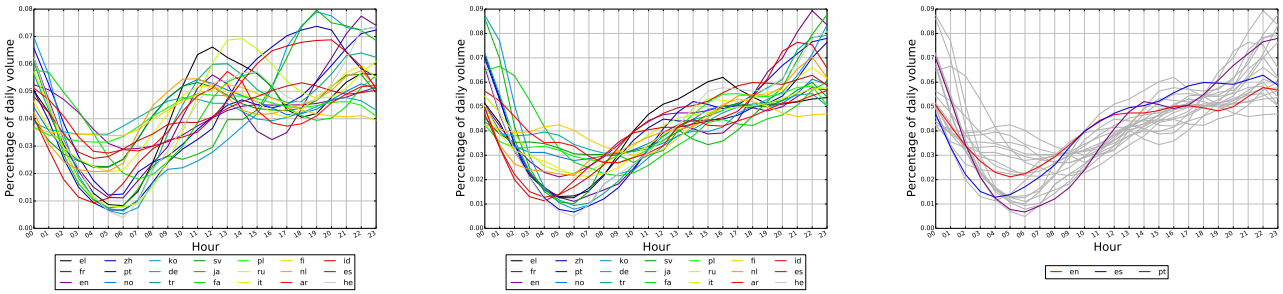
TABLE II: NUMBER OF RECEIVED TWEETS PER LANGUAGE AND PER STREAM FROM FEBRUARY 1ST 2013 TO FEBRUARY 1ST 2014.

Language	Abbreviation	UTC Timezone	Sample	Geo-located	'NL general'	'NL specific'
English	en	-5	457,243,925 (33.85%)	503,979,412	26,590,575	8,805,644
Japanese	ja	9	214,383,682 (15.87%)	39,163,635	1,565,480	72,675
Spanish	es	-5	159,954,649 (11.84%)	145,905,496	3,162,450	476,042
Indonesian	id	7 (Jakarta)	116,797,591 (8.65%)	173,151,505	12,490,099	758,133
Portuguese	pt	-3 (Brasilia)	75,455,200 (5.59%)	137,408,037	1,840,142	250,478
Arabic	ar	2 (Egypt)	72,798,062 (5.39%)	40,181,252	64,144	11,973
Turkish	tr	2	31,035,914 (2.3%)	62,847,302	9,698,937	97,305
French	fr	1	28,284,488 (2.09%)	47,852,027	2,870,243	342,836
Russian	ru	4 (Moskow)	24,798,379 (1.84%)	29,639,926	602,388	22,434
Korean	ko	9	16,506,590 (1.22%)	5,474,115	65,661	10,385
Dutch	nl	1	13,270,846 (0.98%)	16,481,387	567,200,368	110,396,915
Italian	it	1	11,145,933 (0.83%)	14,450,810	316,434	79,364
German	de	1	8,919,771 (0.66%)	9,808,823	12,100,125	898,117
Polish	pl	1	6,044,235 (0.45%)	6,904,949	1,004,977	203,577
Swedish	sv	1	3,347,244 (0.25%)	6,247,073	2,131,957	168,306
Finnish	fi	2	1,687,673 (0.11%)	2,322,875	733,630	84,428
Greek	el	2	1,075,685 (0.08%)	921,053	18,737	770
Persian (Farsi)	fa	4	1,035,230 (0.08%)	746,396	2,537	1,081
Norwegian	no	1	870,515 (0.06%)	1,414,377	528,787	193,297
Chinese	zh	8	809,889 (0.06%)	1,188,903	14,126	2,101
Hebrew	he	2	460,390 (0.03%)	1,222,151	2,042	152
Other			104,715,420 (7,75%)	122,068,237	9,814,957	1,383,383
Total			1,350,641,311	1,369,379,741	652,818,796	124,259,396



(a) max #tweets $\geq 200,000$ (b) $10,000 \leq$ max #tweets $< 200,000$ (c) max #tweets $< 10,000$

Figure 2: Yearly patterns for tweet volume in sample stream.

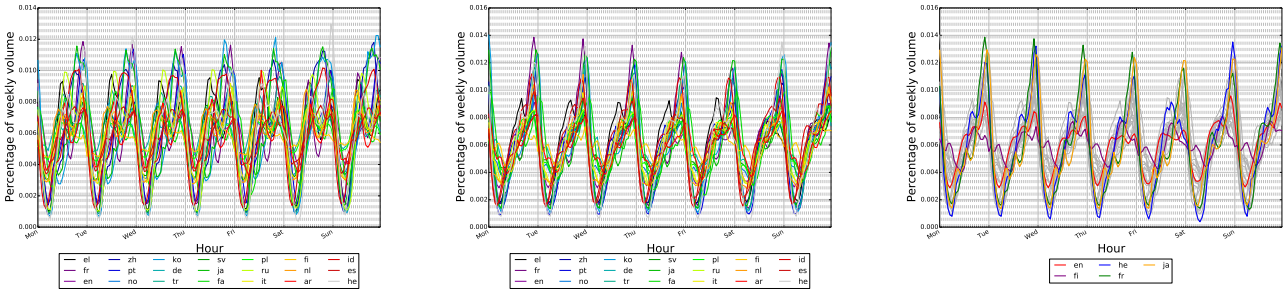


(a) Daily cycle geo-located stream (b) Daily cycle sample stream (c) Daily cycle sample stream for English, Spanish, and Portuguese

Figure 4: Average daily tweet patterns.

First, we analyse the trend in the daily number of tweets for the sample stream. We see large differences between languages. Since several plots greatly overlap, we display these plots in Figure 2 in three separate figures. The clear majority of the tweets we found are written in English (see Figure 2a), followed by the number of tweets in Japanese. Next we find Spanish, Indonesian, Portuguese, and Arabic. In Figure 2b, we select all languages that do not have more than 200,000 tweets per day in our dataset and in Figure 2c we choose this number to be 10,000. During February 2013, all plots in Figure 2

 ISBN: 978-1-61208-358-2



(a) Weekly cycle geo-located stream

(b) Weekly cycle sample stream

(c) Weekly cycle sample stream for English, Hebrew, Japanese, Finnish, and French

Figure 5: Average weekly tweet patterns.

indicate 0 tweets per language. This is because Twitter updated the language detection algorithms during this time [21]. After language detection is turned on, there are still clear jumps for certain languages. Most likely, these jumps correlate with updates of language detection algorithms in these languages. For the number of Dutch tweets in the sample stream, we see that this quantity is decreasing. We note that a decreasing line in Figure 2 does not have to imply that the number of tweets of that language is decreasing, since the quantities in the sample stream are relative to the total Twitter stream. However, Figure 3 shows the daily number of Dutch tweets that were received for all four streams. We see that all four streams

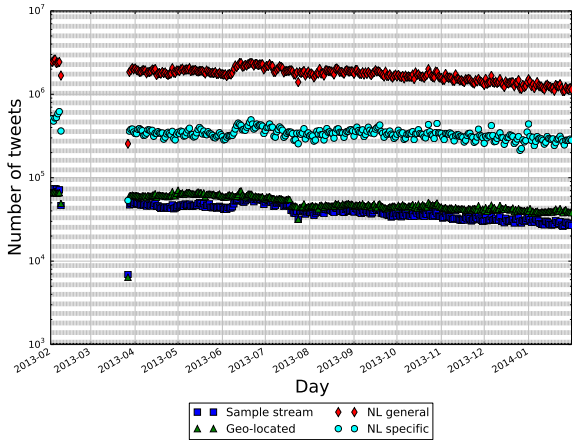


Figure 3: Number of Dutch tweets scraped daily.

show a decreasing pattern. Thus, we can conclude that the total volume of Dutch tweets is decreasing. This fact is also supported by Figure 1, in which the number of missed tweets for the 'NL general' stream is decreasing.

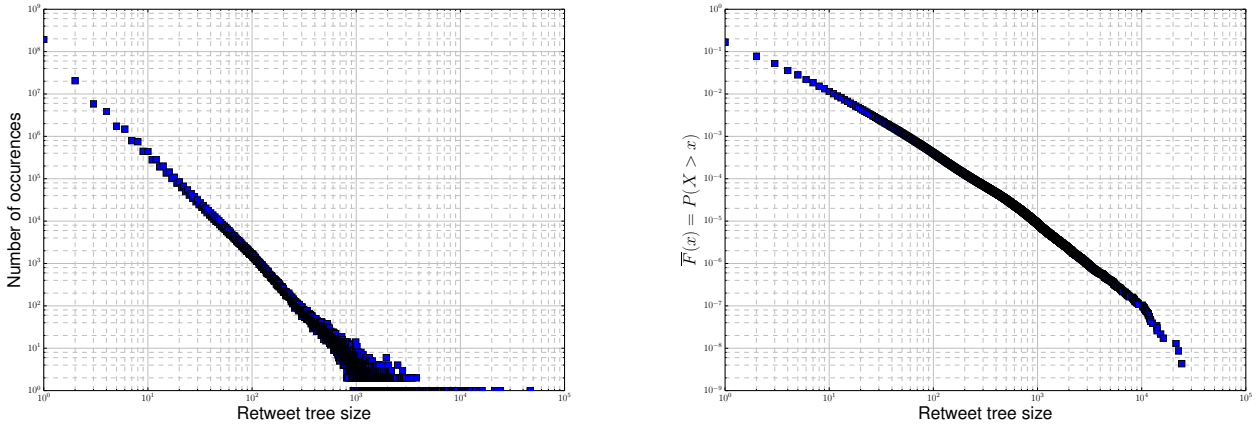
Second, we focus on the daily patterns in the tweet volume. Since all languages are spoken in different timezones, we adjust the time-series to UTC time. The corrections that we used can be found in Table II. We analyse the daily patterns of two streams, namely the geo-located and the sample stream (Figures 4a and 4b, respectively). One striking difference between these figures is that while the sample stream displays a similar pattern for all languages, the geo-located stream differs strongly between languages. In the geo-located stream, we see two clear intervals where the activity peaks, namely during lunch hours and during the evening, which concurs with the findings of [9]. In Figure 4c, we highlight the daily patterns of three languages: English, Spanish, and Portuguese. These patterns are rescaled to American times, thus the majority of the usage of Twitter in these languages originates there. Furthermore, the amplitude of the English pattern is very low, therefore the activity in English is very spread out during the day. A large contrast to this is the pattern in Portuguese tweets. In this pattern, we clearly see a large decay in the number of tweets during the hours of the night.

Third, we regard weekly patterns in the tweet volume. Again, we focus on the geo-located and the sample stream. Similar to the daily patterns, we see that the sample stream patterns in Figure 5b are more consistent than the patterns in the geo-located streams in Figure 5a. For the majority of languages, we see a decreasing activity throughout the week in the sample stream patterns. Two good examples of this 'standard' pattern are Japanese and French, which are highlighted in Figure 5c. After a weekly decrease, the activity increases again on Sunday. Another language that follows this pattern is English; however, it has a very low amplitude with respect to the aforementioned languages. However, a clear exception to this pattern is the Hebrew pattern. In this language, the increase in activity happens on Saturday (since it is the Sabbat). Further, we find that for languages for which we have a small number of tweets (e.g., Finnish), there is only a decrease in activity during the night. The afternoon and evening activity for these languages are evenly distributed.

V. RETWEET TREES

In this section, we analyse the progression of tweets in our dataset through retweets. We define the progression of a message m as the retweet tree related to that message, which we denote by T_m. The graph T_m is a rooted tree, where all non-root vertices indicate people who retweeted the original message. If a user retweets an already retweeted message, this is shown as a new level in T_m. We use all four streams to determine the retweet trees. Figure 8 gives some examples of retweet trees.

 ISBN: 978-1-61208-358-2



(a) Distribution of the retweet tree size

(b) CCDF of the retweet tree size

Figure 6: Retweet tree size.

Require: $(t_{TS}, t_{MID}, t_{UID})\ \forall\ t \in V_{T_m}$ and root node $(r_{TS}, r_{MID}, r_{UID})$.

1: stop = False; $C = \{r_{MID}\}$;$L_c = \{(r_{MID}, r_{UID})\}$; $L_n = \emptyset$.
2: **while** stop = False **do**
3: $T = V_{T_m} \setminus C$;
4: $T^* = sort\_on\_time(T)$ {Time of posting}
5: **for** $(t_{MID}, t_{UID}) \in T^*$ **do**
6: **for** $(u_{MID}, u_{UID}) \in L_c$: **do**
7: **if** t_{UID} is followed by u_{UID}: **then**
8: $(t_{MID}, u_{MID}) \rightarrow E_{T_m}$
9: $u_{MID} \rightarrow C$
10: $u_{MID} \rightarrow L_n$
11: **end if**
12: **end for**
13: **end for**
14: **if** $L_n = \emptyset$ **then**
15: stop = True
16: **else**
17: $L_n \rightarrow L_c$
18: $L_n = \emptyset$
19: **end if**
20: **end while**
21: **for** $t \in V_{T_m}$: **do**
22: **if** $t_{MID} \notin E_{T_m}$ **then**
23: $(r_{MID}, t_{MID}) \rightarrow E_{T_m}$
24: **end if**
25: **end for**
26: **return** E_{T_m}

Figure 7: Determine progression of message m.

We aim to derive the distribution function for the size of a retweet tree. Thus for each retweet tree in our dataset, we calculate the number of nodes in T_m, denoted by $|V_{T_m}|$, and use this to build the distribution function. This function is displayed in Figure 6a. Using this distribution function, we determine the Complementary Cumulative Density Function (CCDF) of the retweet tree size (see Figure 6b). From these plots, we find that the retweet tree size follows a power-law distribution.

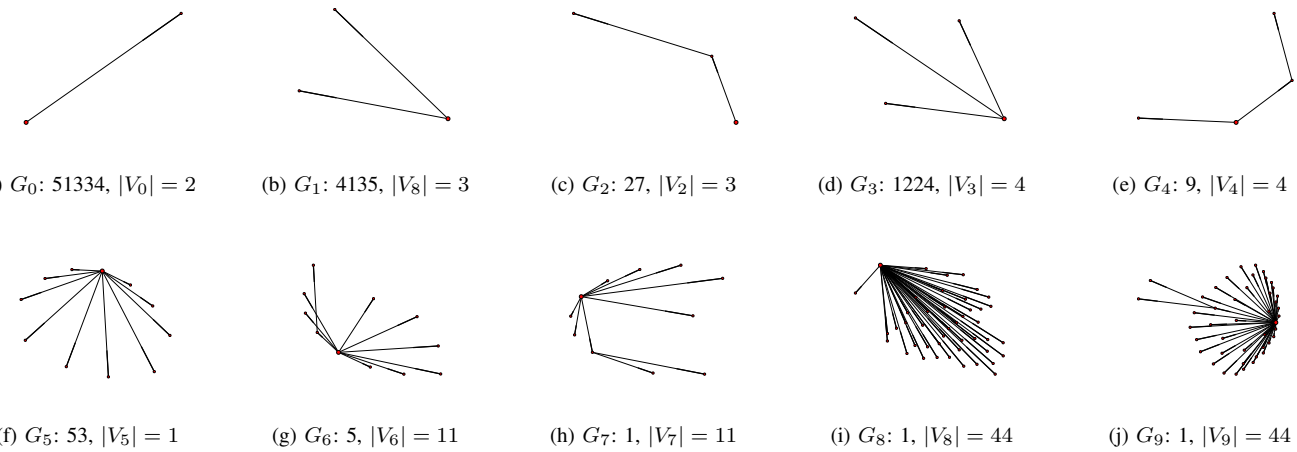
When a retweet is received in the Twitter API, one also receives the original message. However, the level at which this message lies in the retweet tree T_m is not given. Therefore, we propose the following algorithm to determine the progression of a retweet tree T_m. Given all retweets of a certain message, we start with this message. Then, for all retweets, we find out if the user that made that retweet is following the original poster of the message. If this is the case, this user retweeted the original poster. After we checked all users, we find the first level of the retweet tree. If we iterate this procedure until we cannot add new retweets to the tree, we are done. However, using this approach, it could be the case that some retweets have not been placed in the tree. Since these users do not follow any of the other users, we assume they found the tweet through search and thus retweet the original message. This algorithm is defined more formally in Figure 7. Here, we denote L_c as the current level, L_n as the new level and C is a list of checked messages. Furthermore, TS is short for timestamp, MID and UID are the message and user ID number, respectively, and E_{T_m} indicates the edge-set of tree T_m.

Hereafter, we studied the progression of retweet trees for January 13th 2014 from 18h to 19h. We chose a smaller dataset for this analysis due to the rate limitation of the Twitter API, which we need to check the follow-relation between two users in line 7 of Figure 7.

After retrieving the follow-relations and after processing the retweets, we find that the retweet trees tend to be wide and shallow. For instance the retweet tree consisting of three nodes of a star-shape (Figure 8b) occurs 4,135 times whereas the path-shape retweet tree of three nodes (Figure 8c) only occurs 27 times. Although part of this preference is caused by the last part of our algorithm, we find that we can allocate 67.71% of the 93,579 retweets of the timeframe by using the follow-relations in Figure 7.

VI. Conclusion and Future Work

In this paper, we analysed two aspects of user behaviour in Twitter. First, we analysed daily and weekly patterns that emerge from user activity in Twitter. We found that there



(a) G_0: 51334, $|V_0| = 2$ (b) G_1: 4135, $|V_8| = 3$ (c) G_2: 27, $|V_2| = 3$ (d) G_3: 1224, $|V_3| = 4$ (e) G_4: 9, $|V_4| = 4$

(f) G_5: 53, $|V_5| = 1$ (g) G_6: 5, $|V_6| = 11$ (h) G_7: 1, $|V_7| = 11$ (i) G_8: 1, $|V_8| = 44$ (j) G_9: 1, $|V_9| = 44$

Figure 8: Examples of retweet trees: number of occurences, retweet tree size.

are clear circadian patterns for every language we studied. Moreover, all studied languages show a similar daily pattern throughout the week. This concurs with studies done for other social media.

Also, we examined the number of daily tweets per language. Here we found no global patterns that hold for every language. However, through an analysis of the number of tweets that were not received through the streaming API, we see that the percentage of tweets that contains geo-locational data is increasing over time.

Furthermore, we studied the distribution of the size of a retweet tree. We found that these sizes follow a power-law distribution. Moreover, we extended this analysis to the actual progression of retweet trees in Twitter and found that retweet trees tend to be wide and shallow in their structure. For the algorithm in Figure 7, we need to know the network of relations within Twitter. Since this is a very time-consuming process, a possible extension of this work is to find a way to determine the progression of a retweet tree without knowing the complete graph.

REFERENCES

[1] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, pp. 178–185, 2010.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 492–499.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World Wide Web*. ACM, 2010, pp. 851–860.

[5] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López, "Homogeneous temporal activity patterns in a large online communication space," *IADIS International Journal on WWW/INTERNET*, vol. 6, no. 1, pp. 61–76, 2008.

[6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. New York, NY, USA: ACM, 2007, pp. 15–28.

[7] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[8] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in Foursquare." *ICWSM*, vol. 11, pp. 70–573, 2011.

[9] N. Grinberg, M. Naaman, B. Shaw, and G. Lotan, "Extracting diurnal patterns of real world activity from social media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM–13)*, 2013.

[10] T. Yasseri, R. Sumi, and J. Kertész, "Circadian patterns of Wikipedia editorial activity: A demographic analysis," *PloS one*, vol. 7, no. 1, p. e30091, 2012.

[11] M. ten Thij, Y. Volkovich, D. Laniado, and A. Kaltenbrunner, "Modeling and predicting page-view dynamics on Wikipedia," *arXiv preprint arXiv:1212.5943*, 2012.

[12] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes, "Do all birds tweet the same?: characterizing Twitter around the world," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1025–1030.

[13] D. Mocanu, A. Baronchelli, N. Perra, B. Gonalves, Q. Zhang, and A. Vespignani, "The Twitter of babel: Mapping world languages through microblogging platforms," *PLoS ONE*, vol. 8, no. 4, 2013.

[14] D. Jurgens and T.-C. Lu, "Temporal motifs reveal the dynamics of editor interactions in wikipedia," in *International AAAI Conference on Weblogs and Social Media*, 2012.

[15] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks." *ICWSM*, vol. 10, pp. 90–97, 2010.

[16] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng, "Spatio-temporal dynamics of online memes: A study of geo-tagged tweets," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 667–678.

[17] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 177–186.

[18] S. Bhamidi, J. M. Steele, and T. Zaman, "Twitter event networks and the superstar model," *arXiv preprint arXiv:1211.3090*, 2012.

[19] Retrieved: June 27, 2014. [Online]. Available: https://dev.twitter.com/docs/api/1.1/post/statuses/filter

[20] Retrieved: June 27, 2014. [Online]. Available: https://dev.twitter.com/docs/faq#6861

[21] Retrieved: June 27, 2014. [Online]. Available: https://blog.twitter.com/2013/introducing-new-metadata-for-tweets

 ISBN: 978-1-61208-358-2

Enhancement of Trajectory Ontology Inference Over Domain and Temporal Rules

Rouaa Wannous, Jamal Malki and Alain Bouju
L3i laboratory
University of La Rochelle
La Rochelle, France
Emails: {rwanno01, jmalki, abouju}@univ-lr.fr

Cecile Vincent
UMR 7372, CNRS
University of La Rochelle
La Rochelle, France
Email: cvincent@univ-lr.fr

***Abstract*—Capture devices rise large scale trajectory data from moving objects. These devices use different technologies like global navigation satellite system (GNSS), wireless communication, radio-frequency identification (RFID), and other sensors. Huge trajectory data are available today. In this paper, we use an ontological data modeling approach to build a trajectory ontology from such large data. This ontology contains temporal concepts, so we map it to a temporal ontology. We present an implementation framework for declarative and imperative parts of ontology rules in a semantic data store. An inference mechanism is computed over these semantic data. The computational time and memory of the inference increases very rapidly as a function of the data size. For this reason, we propose a two-tier inference filters on data. The primary filter analyzes the trajectory data considering all the possible domain constraints. The analyzed data are considered as the first knowledge base. The secondary filter then computes the inference over the filtered trajectory data and yields to the final knowledge base, that the user can query.**

Keywords–Trajectory ontology modeling; Ontology inference; Domain rules; Temporal rules; Data filter algorithm.

I. INTRODUCTION

Advances in information and communication technologies have encouraged collecting spatial, temporal and spatio-temporal data of moving objects [1]. The raw data captured, commonly called trajectories, traces moving objects from a departure point to a destination point as sequences of data (sample points captured, time of the capture). Raw trajectories do not contain goals of traveling nor activities accomplished by the moving object. Large datasets need to be analyzed and modeled to tackle user's requirements. To answer user's queries we also need to take into account the domain knowledge.

This paper deals with marine mammals tracking applications, namely seal trajectories. Trajectory data are captured by sensors included in a tag glued to the fur of the animal behind the head. The captured trajectories consist of spatial, temporal and spatio-temporal data. Trajectories data can also contain some meta-data. These datasets are organized into sequences. Every sequence, mapped to a temporal interval, characterizes a defined state of the animal. In our application, we consider three main states of a seal: *hauling out, diving* and *cruising*. Every state is related to a seal's activity. For example, a foraging activity of seal occurs during the state diving.

Our goal is to enrich trajectory data with semantics to extract more knowledge. In our previous work [2], we tackled trajectory data connected to other temporal and spatial sources of information. We directly computed the inference over these data. The experimental results addressed the running time and memory problems over the ontology inference computation. Furthermore, we tried to solve these problems by defining some domain constraints, time restrictions in [3] and inference refinements in [4]. The proposed refinements enhanced the inference computation, however, they did not fully solve the problems.

In the present work, we introduce two-tier inference filters on trajectory data. In other words, two distinct operations are performed to enhance the inference: primary and secondary filter operations. The primary filter is applied to the captured data with the consideration of domain constraints. The primary filter allows fast selection of the analyzed data to pass along to the secondary filter. The latter computes the inference over the data output of the primary filter. The global view of this work is detailed as the following steps:

- Semantic trajectory data is an RDF dataset based on an ontology trajectory;
- For analyzing the data, filtering or indexing could be applied. In our case, we carry out a place-of-interest process to analyze data. The analyzed data are stored in a knowledge repository;
- The secondary filter computes inferences over the data with the consideration of domain knowledge;
- The semantic trajectory data and the new data inferred are stored in the knowledge repository.

This paper is organized as follows. Section II summarizes recent work related to trajectory data modeling using ontology approach and some introduced solutions to tackle the problem of the inference complexity using data filtering. Section III illustrates an overview of the ontological modeling approach used. This trajectory ontology contains temporal concepts mapped to W3C OWL-Time ontology [5] in Section IV. Section V details the implementation of the trajectory ontology, the domain ontology rules and the temporal rules. Section VI addresses the complexity of the ontology inference over the domain and temporal rules. Section VII introduces the primary filter over trajectory data based on a place-of-interest process. Section VIII evaluates the ontology inference over the filtered data. Finally, Section IX concludes this paper and presents some prospects.

II. RELATED WORK

Data management techniques including modeling, indexing, inferencing and querying large data have been actively investigated during the last decade [4][6][7]. Most of these techniques are only interested in representing and querying moving object trajectories [2][4][8]. A conceptual view on trajectories is proposed by Spaccapietra et al. [9] in which trajectories are a set of stops, moves. Each part contains a set of

 ISBN: 978-1-61208-358-2

semantic data. Based on this conceptual model, several studies have been proposed, such as [8][10]. Yan et al. [8] proposed a trajectory computing platform that exploits a spatio-semantic trajectory model. One of the layers of this platform is a data preprocessing layer which cleanses the raw GPS feed, in terms of preliminary tasks such as outliers removal and regression-based smoothing. Alvares et al. [10] proposed a trajectory data preprocessing method to integrate trajectories with the space. Their application concerned daily trips of employees from home to work and back. However, the scope of their paper is limited to the formal definition of semantic trajectories with the space and time without any implementation and evaluation. Trajectory filtering visualises a subset of available trajectories [11]. This is useful to view interesting trajectories and discard uninteresting ones. Trajectory filtering can be run in two modes: soft, hard filtering.

Based on a space-time ontology and events approach, Boulmakoul et al. [12] proposed a generic meta-model for trajectories to allow independent applications. They processed trajectories data benefit from a high level of interoperability, information sharing. Their approach is inspired by ontologies, however the proposed resulting system is a pure database approach. Boulmakoul et al. have elaborated a meta-model to represent moving objects using a mapping ontology for locations. In extracting information from the instantiated model during the evaluation phase, they seem to rely on a pure SQL-based approach not on semantic queries. Taking these limitations into account, we defined and implemented two tier inference filters over trajectory data to clean and analyze the data and solve the inference computation problem. Baglioni et al. [13][14] are based on the conceptual model on trajectories [9]. They represent annotated trajectories in an ontology encompassing geographical and application domain knowledge. They consider different kinds of stops and temporal knowledge to discriminate among them. Afterwards, they use ontology axioms to infer behavior of patterns using Oracle and OWLPrime to test the axioms. Moreover, Perry et al. in [15] apply an inference mechanism over their ontology. This inference is based on several domain specific table functions and only on RDFS rules indexes. They use a military application domain and apply complex queries require sophisticated inference methods. In their implementation, they use Oracle DBMS and demonstrate the scalability of their approach with a performance study using both synthetic and real-world RDF datasets.

III. TRAJECTORY ONTOLOGY MODELING

A. Trajectory Domain Ontology

This paper considers trajectories of seals. The data are provided by LIENSs [16] laboratory in collaboration with SMRU [17]. These laboratories work on marine mammals' ecology. Trajectory data of seals between their haulout sites along the coasts of the English Channel or in the Celtic and Irish seas are captured using GNSS systems.

From the analysis of the captured data, we define a seal trajectory ontology that we connect to the trajectory domain ontology. The trajectory domain ontology is our model used in many moving object applications. Details of the modeling approach is discussed by Mefteh [18]. Figure 1 shows an extract of the seal trajectory ontology, called `owlSealTrajectory`. Table I gives a dictionary of its concepts and their relationships.

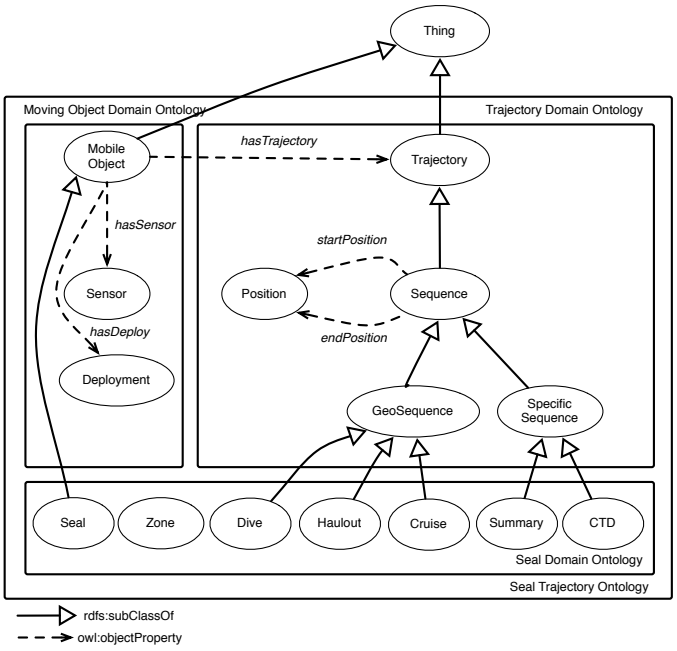


Figure 1. Overview of the seal trajectory ontology

B. Seal Trajectory Ontology

In this work, we propose a Semantic Domain Ontology (Figure 2) based on activities organized as general ones linked to trajectory, and a hierarchy of basic activities linked to sequences of the trajectory domain ontology.

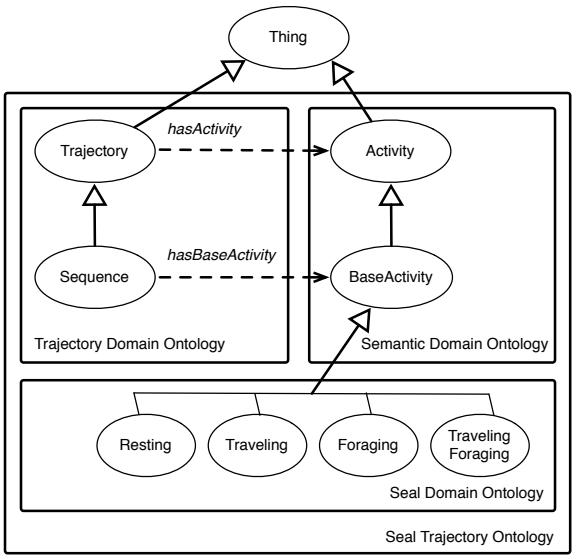


Figure 2. Overview of Seal Trajectory Ontology

The Seal Domain Ontology (Figure 2) is dealing with seal's activities. According to the domain expert, four activities (*resting, traveling, foraging* and *traveling-foraging*) are related to the three states of a seal. The seal trajectory ontology sequences are associated with these main activities.

 ISBN: 978-1-61208-358-2

Table I. Seal Trajectory Ontology Dictionary

Trajectory domain ontology	
Concept	**Description**
Trajectory	logical form to represent sets of sequences
Sequence	spatio-temporal interval representing a capture
GeoSequence	spatial part of sequence
Specific Sequence	metadata associated of a capture
startPosition, endPosition	object properties to represent the end and the beginning of a sequence
Seal domain ontology	
Concept	**Description**
haulout	a state of a seal when it is out of the water (on land) for at least 10 minutes
cruise	a state of a seal where it is in the water and shallower than 1.5 meter
dive	a state of a seal where it is in the water and deeper than 1.5 m for 8 seconds
Summary, CTD	metadata about deployment's conditions of the sensor, marine environment
dive_dur, sur_dur, max_depth	data properties: dive duration, surface duration and maximum depth of a dive, respectively
TAD	Time Allocation at Depth: data properties to define the shape of a seal's dive [5]

IV. Time Ontology

The seal trajectory ontology includes concepts that can be considered as temporal. For example, the concept `Sequence` is a temporal interval. To integrate temporal concepts and relationships in the seal trajectory ontology, we choose a mapping approach between our ontology and the `OWL-Time` ontology [5] developed by the World Wide Web Consortium (W3C). This mapping is detailed in our previous work [2]. An extract of the declarative part of this ontology is shown in Figure 3 described in detail by Jerry and Feng [5].

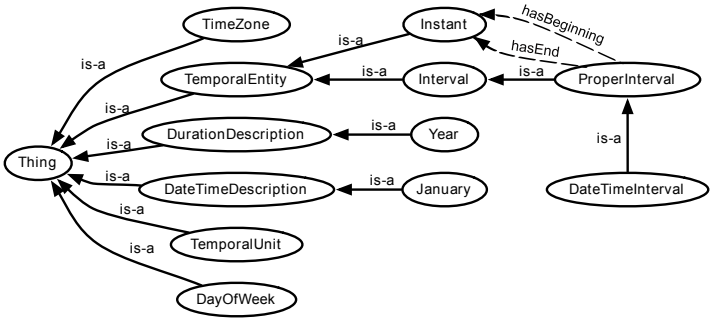


Figure 3. A view of the `OWL-Time` ontology

We are mainly interested in the `ProperInterval` concept and its two properties `hasBeginning` and `hasEnd`.

V. Implementation of Ontologies

A. General Framework Implementation

For the implementation of the ontologies, we use Oracle Semantic Technologies. These technologies have evolved since Oracle DBMS version 10g, 11g and take the name of "Oracle Spatial and Graph - RDF Semantic Graph" in Oracle DBMS version 12c. This system provides support for persistence, inference and querying ontologies through implementation of RDF, RDFS and a large part of OWL standards. The DBMS defines a core in its metabase to support technologies related to ontological data. It stores the ontology declaration with data as RDF triples in the system under the scheme `MDSYS`. Each triple {subject, predicate, object} is handled as a basic data object. Detailed description of this technology can be found in Oracle Semantic Technologies Developer's Guide [19]. To create declarative and imperative parts of the seal trajectory and time ontologies, we:

1) Create the declarative parts of the ontologies;
2) Create instances and population of the ontologies;
3) Consistency checking of the ontological instances;
4) Create the imperative parts of the ontology (seal trajectory ontology rules and temporal rules).

B. Seal Trajectory Ontology Rules

The seal trajectory ontology (Figure 2) is dealing with the seal's activities. Each seal activity has both a declarative part and an imperative corresponding part. The imperative parts of the activities are defined as rules in the ontology. A rule is an object that can be used by an inference process to query semantic data.

Oracle Semantic Technologies is a rule-based system where rules are based on `IF-THEN` patterns and new assertions are placed into working memory. Thus, the rule-based system is said to be a deduction system. In deduction systems, the convention is to refer to each `IF` pattern an antecedent and to each `THEN` pattern a consequent. User-defined rules are defined using the `SEM_APIS.CREATE_RULEBASE` procedure in a `rulebase`. Our rulebase is called `sealActivities_rb`. The system automatically associates a view called `MDSYS.SEMR_rulebase-name` to insert, delete or modify rules in a `rulebase`. Figure 4 gives the `foraging_rule` definition based on domain expert's conditions. From line 4 to 10 of Figure 4, we construct a subgraph and necessary variables needed by the `IF` part of the `foraging_rule`. Line 11 gives the `THEN` part of the rule. Line 12 defines the namespace of ontology.

```
1  EXECUTE SEM_APIS.CREATE_RULEBASE('sealActivities_rb');
2  INSERT INTO mdsys.semr_sealActivities_rb
3  VALUES( 'foraging_rule',
4  '(?diveObject rdf:type          s:Dive             )
5   (?diveObject s:max_depth       ?maxDepth          )
6   (?diveObject s:tad             ?diveTAD           )
7   (?diveObject s:dive_dur        ?diveDur           )
8   (?diveObject s:surf_dur        ?surfaceDur        )
9   (?diveObject s:seqHasActivity  ?activityProberty )',
10 '(maxDepth > 3) and (diveTAD > 0.9) and
            (surfaceDur/diveDur < 0.5)',
11 '(?activityProberty rdf:type  s:Foraging          )',
12 SEM_ALIASES(SEM_ALIAS('s','owlSealTrajectory#')));
```

Figure 4. Implementation of the `foraging rule`

C. Time Ontology Rules

The OWL-Time ontology declares the 13 temporal interval relationships based on Allen algebra [20]. We implement the rule base owlTime_rb to hold the interval temporal relationships. For example, Figure 5 presents the implementation of the imperative part of the intervalAfter_rule based on operations defined in the table TM_RelativePosition of the ISO/TC 211 specification about the temporal schema [21].

```
1  EXECUTE SEM_APIS.CREATE_RULEBASE('owlTime_rb')
2  INSERT INTO mdsys.semr_owltime_rb
3  VALUES('intervalAfter_rule',
4  '(?tObj1    rdf:type ot:ProperInterval        )
5   (?tObj2    rdf:type owltime:ProperInterval   )
6   (?tObj1    ot:hasEnd ?end1                   )
7   (?end1     :inXSDDateTime ?endTime1          )
8   (?tObj2    ot:hasBeginning ?begin2           )
9   (?begin2   ot:inXSDDateTime ?beginTime2      )',
10 '(beginTime2 > endTime1                       )',
11 '(?tObj2    owltime:intervalAfter ?tObj1      )',
12 SEM_ALIASES(SEM_ALIAS('ot','http://www.w3.org/2006/time#')
       ));
```

Figure 5. Implementation of the intervalAfter rule

In Figure 5, line 10 expresses the condition that the beginning of the reference interval is bigger than the end of the argument interval, as explained in the following condition. Line 11 is the consequent of the rule.

$$self.begin.position > other.end.position$$

where

$$\begin{cases} self & = & tObj2 \\ other & = & tObj1 \\ self.begin.position & = & beginTime2 \\ other.end.position & = & endTime1 \end{cases}$$

VI. Trajectory Ontology Inference

Inferencing is the ability to make logical deductions based on rules defined in the ontology. Inferencing involves the use of rules, either supplied by the reasoner or defined by the user. At the data level, inference is a process of discovering new relationships, in our case, new triples. Inferencing, or computing entailment, is a major contribution of semantic technologies that differentiates them from other technologies.

In Oracle Semantic Technologies, an entailment contains precomputed data inferred from applying a specified set of rulebases to a specified set of semantic models. Figure 6 creates an entailment over the seal trajectory and time models. This entailment uses a subset of OWL rules called OWLPrime [19], the seal trajectory and time ontologies rules. Other options are also required like the number of rounds that the inference engine should run. When applying user-defined rules USER_RULES=T, the number of rounds should be assigned as default to REACH_CLOSURE.

In our experiment, we measure the time needed to compute the entailment (Figure 6) for different sets of real trajectory data for one seal. Its movements are captured from 16 June until 18 July 2011 and we have got 10 000 captured data. In this experiment, the seal activity rulebase contains only the foraging rule. The input data for this entailment are only dives. Figure 7 shows the experiment results for the computation time in seconds needed by the entailment. For example, for

```
1 SEM_APIS.CREATE_ENTAILMENT('owlSealTrajectory_idx',
2 SEM_MODELS('owlSealTrajectory','owlTime'),
3 SEM_RULEBASES('OWLPrime','sealActivities_rb', '
      owlTime_rb'),
4 SEM_APIS.REACH_CLOSURE,
5 NULL,
6 'USER_RULES=T');
```

Figure 6. Entailment over the owlSealTrajectory and owlTime ontologies

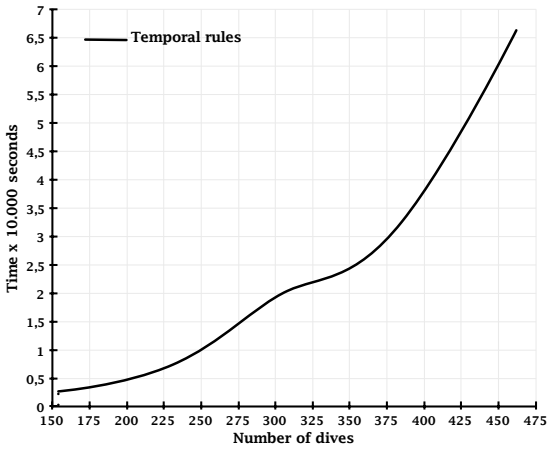


Figure 7. Entailment computation time with all temporal rules and the foraging activity

450 dives, the inference takes around 60 000 seconds ($\simeq$ 16.6 hours).

We notice the huge time taken from the inference mechanism over a small data.

VII. Place Of Interest over Trajectory Data

We introduce a two-tier inference refinement on trajectory data. In other words, two distinct operations are performed to enhance the inference: primary and secondary inference operations. Figure 8 shows the two-tier inference filter refinement. The primary filter is applied to the captured data to classify them into a set of interested places, called Area-Restricted Search (ARSs). The primary filter allows fast selection of the classified data to pass along to the secondary inference. The latter computes the inference mechanism considering the ARS. Then, instead of annotating each sequence in the model, we annotate the ARSs with the expert knowledge activity model. The inference process is computed for each ARS. The secondary inference yields the final knowledge data that the user can query.

Our proposal is to analyze the captured data before computing the ontology inference. This analysis is achieved thanks to our primary filter. This filter considers trajectories that are segmented by the object positions. These positions change and remain fixed. Spaccapietra [9] named the former moves and the latter stops. For this reason, a trajectory is seen as a sequence of moves going from one stop to the next one.

Definition 1 (Stop): A stop is a part of a trajectory having a time interval and represented as a single point.

Definition 2 (Move): A move is a part of a trajectory represented as a spatio-temporal line.

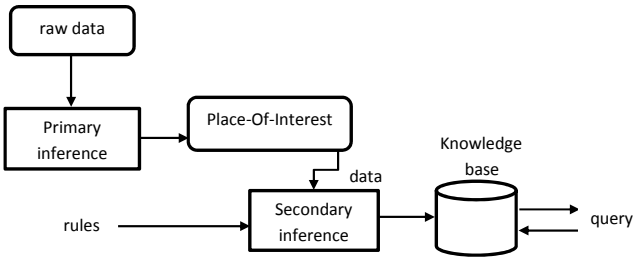


Figure 8. Two-tier inference filter refinement

```
input  : Move
input  : Stop
input  : radius
output: Places
1  initialization;
2  Neighbor ← φ;
3  Points_Neighbors ← φ;
4  Places ← φ;
5  for each p_i ∈ Move do
6      calculate Neighbor(p_i);
7      Points_Neighbors ← (p_i, Neighbors(p_i));
8      Move ← Move − Neighbor(p_i);
9  end
10 for each p_i ∈ Points_Neighbors AND condition(p_i, peaks_i) AND
   condition(distance(p_i, Stop) > radius) do
11     if distance(p_i, Places[j]) > radius then
12         Places[k] ← (Neighbors_i, 1);
13     else
14         Places(Neighbors_j, nVisits_j) = ([Neighbors_j, Neighbors_i], nVisits_j + 1);
15     end
16 end
```

Figure 9. The Place Of Interest algorithm

The primary filter defines interesting places for a moving object. The interesting places are related to where the moving object stays more and visits more often. This filter is explained in Figure 9. This algorithm takes the two parts of a trajectory (move and stop) data as input and gives as output interesting places. The following definitions are used by this algorithm:

Definition 3 (Neighbors): Neighbors for a point (p_i) are a list of points from the Move data where the distance between p_i and any neighbor point is smaller than a fixed radius. $Neighbor(p_i) = \{(p_j)_{j=1}^{n} : p_i, p_j \in Move,\ distance(p_i, p_j) < radius\}$.

Definition 4 (Peak): A peak$_i$ is a cardinality of the list $Neighbor(p_i)$. $(peaks_i)_{i=1}^{n} = \#(Neighbor(p_i))_{i=1}^{n}$.

Definition 5 (Points_Neighbors): Points_Neighbors are a list of points and their neighbors. $Points\_Neighbors = \{(p_i, Neighbors_i)_{i=1}^{n} : p_i, Neighbors_i \in Move\}$.

Definition 6 (Places): Place$_i$ is an interesting place which contains the $Neighbor(p_i)$ and number of its visits ($nVisits$) by the moving object. $Places = \{(Neighbors_i, nVisits_i)_{i=1}^{n} : Neighbors_i \in Move,\ nVisits_i \in number\}$.

The first step of the primary filter, Figure 9 lines 5-9, gathers the move data into groups of neighbors. These groups are defined with respect to a *radius*. This radius is a fixed distance between two points to calculate the neighbors. The candidate of the radius is related to the application view of a trajectory, and is an input for this algorithm. The output of the first step is *Points_Neighbors*, from which the second step starts.

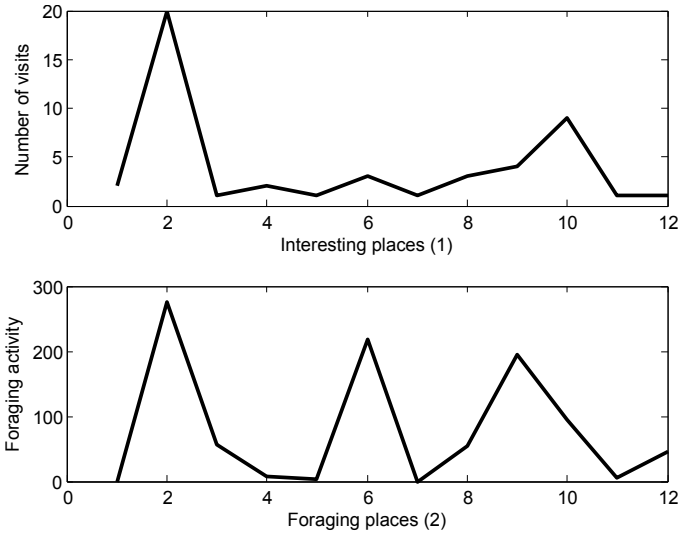


Figure 10. Interesting and foraging places

Lines 10-16, the second step, defines the interesting places. In general, we can consider all the members of *Points_Neighbors* or we can apply a condition over the *Peaks*. For example, the application view could be interesting in places that have 60 points and over, or could be interesting in any place having at least a point. For defining a place, the coordinates of the neighbors could be an interesting place after applying two conditions. Every point that belongs to a place should be far from the stop data more than the fixed radius. Any place should not have any neighbor within the radius distance, otherwise we merge the two coordinates and increase the visits number. The result of this step is the output *Places* of this algorithm.

VIII. Experimental Results

To analyze our data, we consider the same datasets in Section VI. We pass these data to the Place Of Interest algorithm. This algorithm analyzes the data and gives as output the places and their visits, as shown in Figure 10 interesting places (1). However, the main goal is to define foraging places among the captured data from 16 June until 18 July 2011. We look forward to analyse all the 10 000 captured data.

Defining foraging places is the objective of the secondary filter. The secondary filter computes the entailment over the interesting places. This filter specifies foraging places among 10 000 captured data. It determines the number of foraging activity for each place, as shown in Figure 10 foraging places (2). We can notice that the places 1, 4, 5, 7 and 11 are not considered as foraging places. Places 2, 6 , 9 and 10 are the significant foraging places. Finally, the results of the primary filter are decreased the captured data from 10 000 into 6 170 interesting raw trajectories organized in places.

By the normal inference ontology computation results, we could not be able to consider all the captured data. We computed the inference just for 500 raw data. However, using the primary filter and defining the interesting places helped us to define foraging places over all the captured data. These

 ISBN: 978-1-61208-358-2

inferred data are considered as the final knowledge data that the user can query.

IX. CONCLUSION

In this work, we proposed a modeling approach based on ontologies to build a trajectory ontology. Our approach considers three separated ontology models: a general trajectory domain model, a domain knowledge or semantic model and a temporal domain model. We map the spatial concepts in the trajectory ontology to the spatial ontology. To implement the declarative and imperative parts of the ontologies, we consider the framework of Oracle Semantic Data Store. To define the thematic and temporal reasoning, we implement rules related to the considered models. The thematic rules are based on the domain trajectory activities and the temporal rules are based on Allen relationships. Then, we define and apply two-tier inference filters. In other words, two distinct operations are performed to enhance the inference: primary and secondary filter operations. The primary filter analyzes the trajectory data into places of interest. The secondary filter computes the ontology inference over the semantic trajectories using the ontology domain and temporal rules. The latter filters the interesting places into domain activity places. The experimental results show that we are able with the two-tier filters to consider all the captured data, whereas we could not even compute the ontology inference. For the evaluation, we use a PC with Linux system over a processor i5-250M, 2.5GHz and 8G memory.

REFERENCES

[1] R. Güting and M. Schneider, *Moving Objects Databases*. Morgan Kaufmann, 2005.

[2] R. Wannous, J. Malki, A. Bouju, and C. Vincent, "Time integration in semantic trajectories using an ontological modelling approach," in *New Trends in Databases and Information Systems*, ser. Advances in Intelligent Systems and Computing. Springer Berlin Heidelberg, 2013, pp. 187–198.

[3] J. Malki, R. Wannous, A. Bouju, and C. Vincent, "Temporal reasoning in trajectories using an ontological modelling approach." Control and Cybernetics, 2012, pp. 761–777.

[4] R. Wannous, J. Malki, A. Bouju, and C. Vincent, "Modelling mobile object activities based on trajectory ontology rules considering spatial relationship rules," in *Modeling Approaches and Algorithms for Advanced Computer Applications*, ser. Studies in Computational Intelligence. Springer International Publishing, 2013, pp. 249–258.

[5] R. H. Jerry and P. Feng, "An ontology of time for the semantic web," in *ACM Transactions on Asian Language Information Processing*, 2004, pp. 66–85, http://www.w3.org/2006/time.

[6] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, "SeMiTri: A framework for semantic annotation of heterogeneous trajectories," in *Proceedings of the 14th International Conference on Extending Database Technology*. ACM, 2011, pp. 259–270.

[7] J. Malki, A. Bouju, and W. Mefteh, "An ontological approach modeling and reasoning on trajectories. taking into account thematic, temporal and spatial rules," in *TSI. Technique et Science Informatiques*, 2012, pp. 71–96.

[8] Z. Yan, C. Parent, S. Spaccapietra, and D. Chakraborty, "A hybrid model and computing platform for spatio-semantic trajectories," in *The Semantic Web: Research and Applications*. Springer Berlin/Heidelberg, 2010, pp. 60–75.

[9] S. Spaccapietra, C. Parent, M. Damiani, J. Demacedo, F. Porto, and C. Vangenot, "A conceptual view on trajectories," in *Including Special Section: Privacy Aspects of Data Mining Workshop*, 2008, pp. 126–146.

[10] L. O. Alvares and et al., "A model for enriching trajectories with semantic geographical information," in *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*. ACM, 2007, pp. 1–22.

[11] Geant4, *Geant4 User's Guide for Application Developers*, 2011, ch. Visualization/Trajectory Filtering.

[12] A. Boulmakoul, L. Karim, and A. Lbath, "Moving object trajectories meta-model and spatio-temporal queries," in *International Journal of Database Management Systems (IJDMS)*, 2012, pp. 35–54.

[13] M. Baglioni, J. Macedo, C. Renso, and M. Wachowicz, "An ontology-based approach for the semantic modelling and reasoning on trajectories," in *Advances in Conceptual Modeling - Challenges and Opportunities*. Springer Berlin/Heidelberg, 2008, pp. 344–353.

[14] M. Baglioni, J. A. Fernandes de Macedo, C. Renso, R. Trasarti, and M. Wachowicz, "Towards semantic interpretation of movement behavior," in *Advances in GIScience*, ser. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, 2009, pp. 271–288.

[15] P. Matthew, "A framework to support spatial, temporal and thematic analytics over semantic web data," Ph.D. dissertation, Wright State University, 2008.

[16] "LIENSs: Cnrs/university of la rochelle," http://lienss.univ-larochelle.fr/.

[17] "SMRU; sea mammal research unit," http://www.smru.st-and.ac.uk/.

[18] W. Mefteh, "Ontology modeling approach on trajectory over thematic, temporal and spatial domains," Ph.D. dissertation, La Rochelle university, 2013.

[19] Oracle, "Oracle Database Semantic Technologies Developer's guide 11g release 2," 2012. [Online]. Available: http://www.oracle.com/technology/tech/semantic-technologies

[20] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, pp. 832–843, 1983.

[21] ISO/TC_211, "Geographic information – temporal schema, ISO 19108," 2002.

 ISBN: 978-1-61208-358-2

An Annotation Process for Data Visualization Techniques

Geraldo Franciscani Jr., Rodrygo L. T. Santos, Raphael Ottoni, João Paulo Pesce,
Wagner Meira Jr. and Raquel Melo-Minardi
Department of Computer Science - DCC
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{gfrancis, rodrygo, rapha, jpesce, meira, raquelcm}@dcc.ufmg.br

***Abstract*—As the area of information visualization grows, a massive amount of visualization techniques has been developed. Consequently, the choice of an appropriate visualization has become more complex, usually resulting in unsatisfactory data analysis. Although there exist models and classifications that could guide the choice of a visualization technique, they are mostly generalist and do not present a clear methodology for evaluation and evolution. In contrast, we propose an annotation process for data visualization techniques based on an initial capability-driven collection of terms and concepts that encompasses visual components of both well established as well as modern visualization techniques. To demonstrate the initial collections expressiveness, we present a qualitative analysis of an experiment with specialist users at annotating visualization techniques from the D^3 (Data-Driven Documents) library. Furthermore, to show the completeness of the collection, we automatically assess its coverage of all published papers from six major international information visualization conferences since 1995. Our results attest the expressiveness of the initial collection and its coverage of over 99% of the analysed literature. Finally, we discuss the limitations and alternatives for semi-automatically evolving the annotation process as new visualization techniques are developed and how the spread of this type of methodology could benefit the information visualization community.**

Keywords–Annotation Process; Data Visualization; Ontologies; Taxonomies.

I. INTRODUCTION

There has been an increasing need to extract relevant information from data and make sense of it in different contexts. At the same time, it is becoming increasingly difficult to identify frequent patterns and exploit large databases. Human abilities of visual perception and cognition come into play as the need to extrapolate textual forms and explore the graphic field become a necessity. As the information visualization area grows, a vast number of visualization techniques are developed. Nonetheless, ordinary users are not prepared to decide which visualization is the most appropriate for the required analysis and tend to express data unsatisfactorily. As a result, the development of strategies and tools to help users choose visualization techniques that can effectively help in data analysis and sense making has become crucial.

It is vital to organize the knowledge of visualization methods and capabilities being produced, with a focus on making visualization development easy, more tangible and effective. We are reaching a juncture of information overload where it has become challenging, even for experts, to cope with the many approaches on visualizing data produced by the academic and design communities. With that in mind, the knowledge being produced by information scientists in the creation of concepts of classification models, taxonomies and ontologies is a straightforward approach.

According to the *Oxford English Dictionary*, a *taxonomy* is a classification, especially in relation to its general laws or principles; that department of science, or of a particular science or subject, which consists in or relates to classification, especially the systemic classification of living organisms. An *ontology* is the science of study of being; this is a department of metaphysics that relates to the being or essence of things, or to being in the abstract. Researchers have been using such a collection of concepts and terms in the biology field since at least 13 years ago, when Gene Ontology was proposed and broadly adopted [1].

From our perspective, the information visualization area requires a unified annotation process that allows its community to annotate or associate terms to both traditional visualization techniques as well as novel techniques being developed. We believe the collection of terms needed by the information visualization field should primarily be able to describe visualization methods in terms of two main elements: visual components and capabilities. Examples of visual components are dimensionality, the objects used in the visual composition, the types of displays and pre-attentive attributes. By capabilities, we mean broader features that encompass the quantitative relationships being described and visual patterns being revealed, as well as the analytical, navigation and interaction techniques that could be used.

According to Gilchrist [2], the definitions of the terms taxonomy and ontology have been subverted and overlap significantly. Previous works focused on the specification of taxonomies, models and ontologies to describe and study the relationships between terms and visualization techniques [3]–[10]. Most importantly, there were attempts to use such classifications and models to generate recommendation systems and to evaluate techniques [11] [12]. Although those works have some important implications in helping users to express data in a more satisfactory way, they do not represent a consensus between specialists and do not address a clear methodology for progressive evaluation and evolvement, regarding the emergence of new techniques and concepts.

In the present work, we describe the methodology used to propose an annotation process for data visualization techniques based on a collection of terms and concepts that covers visual components of visualization techniques and their capabilities. Next, we select a diverse set of visualizations to be annotated with the proposed collection of terms. Note that, here, we import the term annotation from the biology field where it

 ISBN: 978-1-61208-358-2

means the association of terms of the controlled vocabulary with biological objects. We also propose an FP-tree-based [13] algorithm to organize the set of visualizations in a tree where internal nodes are the collection terms and leaves are the visualizations themselves. The tree is a type of visual index that helped us to evaluate both the collection of terms and the selected set of visualization characteristics. We characterize this tree and show how it provides a macro view of the visualization capabilities. Furthermore, we also automatically assess the coverage of our proposed collection of terms in all published papers from six major international information visualization conferences since 1995. Our results attest the expressiveness of the proposed collection and its coverage of over 99% of the analysed literature. In addition, we discuss alternatives for semi-automatically evolving the annotation process as new visualization techniques are developed and finally, how the spread of this type of methodology could benefit the information visualization community.

The remainder of this paper is organized as follows. Section 2 reviews some works related to the development of models and classifications in the information visualization and visual analytics fields. Section 3 describes all the methods, including the proposed annotation process and, in particular, the use of this process, used to build a tree of visualization techniques and their related terms. Section 3 also describes the algorithms we built and used with that purpose as well as the strategy to automatically assess the presence of the terms in the literature. Section 4 presents an evaluative study of the proposed process and discussions about the adopted methods. Section 5 presents the evaluation results. Finally, Section 6 presents our concluding remarks and future directions.

II. Related Work

Many studies focused on the definition of a consistent ontology / taxonomy to categorize visualizations. Our goal is to identify areas already covered by the ontologies / taxonomies existing in the literature and find related examples that serve as basis for our annotation process. Voigt and Polowinski [14] systematically reviewed existing models and classifications, comparing the strengths and weaknesses of each, as well as establishing relationships among them. As a result, the authors specified an initial unified visualization ontology for classification and synthesis of graphical representations. Although it is complete and comprehensive, the authors do not present a methodology for evaluation and evolvement of the concepts presented.

Duke, Brodlie and Duce [15] built an initial skeleton for a vocabulary that would identify the communication between user and system. Concepts and relationships were considered in more restricted areas such as data, tasks and visual representations. In their study, the authors describe how the relationships between published studies may contribute to the construction of this unified ontology and presented, as a major challenge, the consensus among researchers in this area. Although it was an important attempt to organize and categorize existing knowledge, it presents an early version of the vocabulary that would require more specificity to classify a large set of techniques.

Shu, Avis and Rana [9] presented the design of an ontology focused on providing semantics to aid the discovery of visualization services based on the initial concept proposed by Duke, Brodlie and Duce [15]. Their study defined classes mostly for modeling data and visualizations techniques. However, the presented class names were unreadable for users and some concepts were not addressed, such as tasks and interactions.

Shneiderman [16] proposed the Task by Data Type Taxonomy (*TTT*) for information visualizations, dividing the visualization techniques into seven data types (one-, two-, and three-dimensional data, temporal and multi-dimensional data, tree and network data) and seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extracts). The data types characterize the task-domain information objects and are organized by the problems the users are trying to solve. The seven tasks are at a high level of abstraction and represent user interaction with the visualization or data. In 2012, Shneiderman and Jeffrey [4] proposed an update for *TTT* by presenting a taxonomy of interactive dynamics to help users in evaluating and creating visual analysis tools. The taxonomy consists of 12 task types grouped into three high-level categories (1) data and view specification (visualize, filter, sort, and derive); (2) view manipulation (select, navigate, coordinate, and organize); and (3) analysis process and provenance (record, annotate, share, and guide). Although *TTT* was an interesting step towards categorizing and organizing existing visualizations, from the perspective of visualization annotation, it is still too generalist and could benefit from the addition of more detailed and discriminative terms.

Chi [5] presented another taxonomy based on what they called the Data State Reference model. This model divides each technique into four data stages (value, analytical abstraction, visualization abstraction and view) and three types of data transformation operators (data transformation, visualization transformation and visual mapping transformation). Within each data stage, there are four types of operators that do not change the underlying data structures, the within stage operators (within value, within analytical abstraction, within visualization abstraction and within view). Data transformation operators are used to transform data from one stage to another, and within stage operators are used to transform data without changing the underlying data structure. The contribution of this model is in the sense that the authors classified each visualization technique by not only its data type but also its processing operating steps, which helps in understanding the operating steps for each classified visualization technique and in defining sequential ordering of operations and their dependencies. However, this model is limited in comparison to our proposal regarding visualization annotation process in the sense that it does not take into account important factors about the expressive power of visualization techniques in terms of what quantitative relationships they are able to represent, what type of data they can present and what type of visual patterns they can evidence. Additionally, this model does not consider visual objects and pre-attentive attributes involved in the representations.

A different taxonomy-based approach is to focus on the visualization algorithm instead of the data to be visualized. Tory and Möller [6] proposed a model divided into four categories: object of study, data, design model and user model. This model does not attempt to consider the data-oriented approach, instead emphasizing a more flexible system that

 ISBN: 978-1-61208-358-2

highlights the users' conceptual model of the visualization.

Fujishiro, Furuhata, Ichikawa and Takeshima [11] presented a semi-automatic approach for the development of data visualization applications. The authors proposed the GADGET/IV system, based on a goal-oriented taxonomy. This taxonomy has been constructed by combining the Wehrend Matrix [17] with the concepts introduced in TTT [16]. Moreover, this system was an extension to the GADGET (Goal-oriented Application Design Guidance for modular visualization EnvironmenTs) system [18], which used only the above matrix as a reference to aid the development of data visualization applications. This research presented an interesting perspective, although the use of the system was not evaluated.

Pfitzner, Hobbs and Powers [8] built a taxonomy-based framework that encompasses several aspects in information visualization: data, tasks, interactions, context and human capacities of cognition. Although the study seems promising and complete, the usefulness of the taxonomy created was not evaluated and it lacks a clear methodology for evolving the taxonomy with the area.

Gilson, Silva, Grant and Chen [19] proposed an ontology as part of a tool that automatically generates visualizations from web pages in specific areas without prior knowledge of the content of these pages. Although the proposed ontology presents properties of graphical representations and visual objects, some important topics such interactions, tasks to be performed on data and user goals were not considered.

Amar, Eagan and Stasko [20] presented a set of ten low-level analysis tasks (retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster and correlate). According to the authors, these tasks capture people's activities while employing information visualization techniques to understand data. These tasks were obtained using an affinity diagramming approach from 200 sample questions from students about how they would analyze five different datasets from different domains with information visualization tools. Despite being very interesting, this taxonomy focuses only on analytical tasks and not on visualization techniques, which is what this work focuses on.

Zhou and Feiner [10] developed a visual task taxonomy that extends the one proposed by Wehrend and Lewis [17]; additional tasks were defined, parameterized, and grouped in three dimensions (organization, signaling and transformation). These dimensions were composed by types and subtypes where elemental tasks were defined (for instance, associate, cluster, locate, categorize, cluster, distinguish, among others). Morse, Lewis and Olsen [12] showed that this type of taxonomy can be used in the evaluation of visualization techniques. In this research, a methodology is developed to create a set of taxonomy-based tasks for evaluating visualization techniques for information retrieval. According to this research, the taxonomies are very useful for addressing the complexity of the visual tasks.

From our point of view, we will consider all related works to compose our annotation process, as they are an important inheritance in the area. However, in this work, we will not consider data preparation or transformation tasks. We are mainly interested in visual components, which are not considered in most previous works, and the capabilities of the visualization techniques, which have been considered with different perspectives. We tried to conserve important terms regarding data type, but the majority of the terms we kept concerns important analytical interaction techniques that can be applied to the visualizations and consequently can give them important capabilities.

III. Annotation Process

The proposed annotation process consists in a definition of a collection of terms and concepts related to a set of data visualizations techniques to be annotated. To this end, we conduct an experiment with experts who defined an initial selection of terms and concepts in existing literature. We also propose an initial set of data visualization techniques that will serve as a source for the study and may also evolve with the area. Finally, we present the annotation process itself as an association of the techniques with the terms and concepts.

A. Initial Collection of Terms and Concepts

We had two main objectives in proposing the initial collection of terms and concepts: terms should describe visualization techniques concerning their visual components; and terms should encompass the quantitative relationships being described and visual patterns being revealed, as well as the analytical, navigation and interaction techniques that could be used with the visualization.

First, we list all terms and concepts found in the existing models and classifications in the literature presented in Section II. Then, we enrich this set with other terms manually selected from references qualified in our research field [21]–[33]. The first reference used terms we considered useful for the two aforementioned objectives we defined for the annotation process and the second is classical in terms of visual objects. As a result, we obtained a set composed by 101 terms.

In order to adjust this initial set with the proposed objectives, we conduct an experiment with three experts (one professor and two MSc students in Information Visualization) and three data visualization research assistants. Each one evaluated the relevance (yes or no) of each term according to the two previously mentioned objectives. After that, we considered the terms that had 100 % positive reviews (63). The terms with one or more negative evaluations were discussed among the group and evaluated again. Terms with an agreement higher than 80% were considered (11), and the remaining disregarded (27). At the end of the experiment, we obtained a more appropriate initial collection composed by 74 terms.

We present the initial collection below. The following terms present visual objects and attributes that are intuitive and self-explanatory. Thus, we only cite them: *Bars, Boxes, Cells, Circle Section, Lines, Points, Ring Sector, Shape, Trails, Motion, Direction, 2D Spatial Position Representing Quantities, Spatial Grouping Position Representing Categories, Blur, Color Variation, Curvature, Enclosure, Orientation Variation, Shape Variation, Size Variation, Texture Variation, Value Variation, 1D (Dimensional), 2D (Dimensional), 3D (Dimensional), Multidimensional.*

Next, we list and explain the remaining terms:

Correlation: How variables relate to and affect one another.

 ISBN: 978-1-61208-358-2

Deviation: How one or more sets of values deviate from a reference set of values, which can be a target, a forecast, same point in the past, immediately prior period, standard or norm.

Distribution: Examining sets of quantitative values to see how the values are distributed from the lowest to highest or to compare and contrast how multiple sets of values are distributed.

Multivariate: The purpose of multivariate analysis is to identify similarities and differences among items, each characterized by a common set of variables.

Part-to-whole: Used when trying to make sense of a total amount (whole), aggregating them by the parts to see how much each part adds to the whole.

Ranking: Items ranked by value.

Time series: One or a set of time-dependent attributes.

Alternating differences: Differences from one value to the next begin small then shift to large and finally shift back again to small.

Center: Estimation of the middle of the set of values.

Co-variation: When two sets of values relate to one another so that changes in one are reflected by changes in the other, either immediately or later, this is called co-variation.

Cycles: Patterns that repeat at regular intervals, such as daily, weekly, monthly, quarterly, yearly, or seasonally.

Exceptions: Values that fall outside the norm.

Gaps: Empty regions where we would expect to find values.

Increasingly different: Differences from one value to the next decrease.

Non-uniformly different: Differences from one value to the next vary significantly.

Rate of change: The percentage difference between one value and the next.

Shape: Shows where the values are located. If it is a curve, for instance, is it curved or flat? If curved, upward or downward? If curved upward, single or multiple peaked? If single peaked, symmetrical or skewed? Concentrations? Gaps?

Spread: A measure of dispersion, that is, how spread out the values are.

Trend: The overall tendency of a series of values to increase, decrease or remain relatively stable during a particular period of time.

Uniform: All values are roughly the same.

Uniformly different: Differences from one value to the next decrease by roughly the same amount.

Variability: The average degree of change from one point in time to the next throughout a particular span of time.

Directed (Analytical Navigation): Begins with a specific question (perhaps a particular pattern), and then produces the answer.

Exploratory (Analytical Navigation): Begins by simply looking at the data without predetermining what might be found. Then, when something that seems interesting is noticed and questioned, we proceed in a directed fashion to find an answer to that question.

Hierarchical (Analytical Navigation): To navigate through information from a high level view into progressively lower levels along a defined hierarchical structure and back up again.

Accessing details on demand: When details are called up instantly when needed but kept out of the way before they are needed and after they have been read. Select a group or item and obtain details when needed.

Adding variables: Adding one or more attributes.

Aggregating: When we aggregate or disaggregate information, we are not changing the amount of information but rather the level of detail at which it is viewed. We aggregate data to view it at a high level of summarization or generalization; we disaggregate to view it at a lower level of detail.

Annotating: To document objects of the display, adding notes to them.

Bookmarking: To allow users to save automatically particular views, including its filters, sorts, and other features, so they can easily return to them later.

Brushing and linking: To highlight the same subset of data in multiple graphs at the same time.

Comparing: Encompasses comparing (looking for similarities) and contrasting (looking for differences).

Drilling: Involves moving down levels of summarization (and also back up) along a defined hierarchical path.

Filtering: The act of reducing the data we are viewing to a subset of what is currently there.

Focus and context together: When we are focusing on details, the whole does not need to be visible in high resolution, but we need to see where the details are focusing or reside within the bigger picture and how they relate to it.

Highlighting: To cause particular data to stand out without causing all other data to go away.

Re-expressing: When we change the way we delineate quantitative values that we are examining (e.g.: changes of units of measure).

Re-scaling: Changes the scale: linear, quadratic, or logarithmic.

Re-visualizing: Changing the visual representation in some fundamental way, such as switching from one type of graph to another.

Sorting: Sorting from low to high or high to low.

Zooming and panning: When we enlarge the portion of the display that we wish to see more closely.

Clustering items by similarity: Clustering is the process of segmenting data into groups whose items share similar features.

 ISBN: 978-1-61208-358-2

Comparison of individual and cumulative values: Useful when we assess how well things are going by comparing actual values to targets.

Multiple concurrent views and brushing: The visualization of a single dataset from different perspectives concurrently using multiple graphs.

Overlapped time scales: We can strengthen our ability to detect and compare cyclical patterns stretching across multiple cycles in a line graph by displaying each cycle as a separate line and overlapping time scales.

Ranking items by similarity: To order items according to their relative similarity to enhance visual analysis.

Reference lines and regions: Objects used to give context to the analysis making comparisons easy. Reference lines usually represent expected values as well as averages or means.

Trellises and cross-tabs / Small multiples: When we divide the data set we wish to examine into multiple graphs, either because we can't display everything in a single graph without resorting to a 3-D display, which would be difficult to decipher, or because placing all the information in a single graph would make it too cluttered to read. By splitting the data into multiple graphs that appear on the screen at the same time in close proximity to one another, we can examine the data in any one graph more easily, and we can compare values and patterns among graphs with relative ease.

B. Visualization Techniques

The visualization techniques used in the study were collected in December of 2012 from D^3's (*Data-Driven Documents*) [34] web site [35]. This dataset was used due to the extensive and varied set of visualizations techniques made available by D^3's collaborators. We removed examples that were not true visualization techniques and represented only examples of how to use the library. A total of 53 visualization techniques remained.

C. Association Process

As noted previously, the annotation term was borrowed from biology and means to associate terms of an ontology with objects of interest. In our case, the ontology is represented by the initial collection of terms and concepts and the objects of interest by the visualization techniques. The annotation process consisted of using a web form to associate a set of terms with each visualization. It was performed by the same team of experts and research assistants and annotations with more than 80% of agreement were considered. We decided to associate with each visualization not only terms that are readily implemented in the visualizations but also every term that could be easily incorporated into the implementation because our purpose is to annotate visualizations according to their capabilities rather than their implementation. Our goal is to open the system to the scientific community to integrate other researchers' opinions about the current annotations in a way that the process will be more robust and reliable, analogous to what happened in biology.

IV. Evaluation Strategy

In this section, we describe our strategies to evaluate the annotation process and its components. Firstly, to evaluate the expressiveness of the initial collection of terms and the performed annotation procedure, we produce a visual index represented in a tree structure. The nodes are the terms, and the leaves are the techniques. Then, to evaluate the completeness of the proposed collection of terms and concepts, we present a methodology for automatically assessing the terms coverage of all published papers from six major international information visualization conferences since 1995.

A. Expressiveness evaluation

To evaluate the expressiveness of the proposed collection of terms and the performed annotation procedure, we produce a visual index in a tree structure. The tree we produced was based on the classical FP-tree which is commonly used to find frequent patterns [36] and to cluster objects [37] in a parameter-independent way. The FP-tree is an appropriate data structure for representing our data because we would like to build a visual index of visualizations and terms capable of grouping similar visualizations in terms of similar visual components and capabilities (the main objectives of our collection of terms and concepts). Additionally, we would like to distinguish popular (and non-discriminative) terms from specific (and discriminative) ones.

We use a modification of the original FP-tree data structure implemented by Pires et al. [37]. Due to space limitations, we will not explain the FP-tree construction algorithm, which can be found with examples in [36]. Each transaction in the database is represented as a path in the tree, where each node is an attribute and the attributes are organized in non-increasing frequency order from root to leaves. The path length (i.e., the number of attributes per transaction) may vary. The attributes are then sorted by their frequency in the database and inserted so that transactions with attributes in common share a path in the tree. Consequently, globally common attributes are at the highest levels and less frequent attributes are at lower levels. The generated tree structure is shown in Figure 2.

B. Completeness evaluation: automatic assessment of the literature coverage

An important drawback of any proposed collection of concepts is the difficulty to assess its completeness. We automatically assess the terms coverage of all published papers from six major international information visualization conferences: IEEE Symposium on Information Visualization (INFOVIS); IEEE Conference on Visual Analytics Science and Technology; EuroVis / Joint Eurographics - IEEE TCVG Symposium on Visualization; International Conference on Information Visualization; Asia Pacific Symposium on information visualization; and Computer Graphics, Imaging and Vision. We download all available papers since 1995, totaling 5,061 publications. To normalize the comparison between the terms and the full-text extracted from each paper, we preprocess all text content by applying standard text processing techniques, such as punctuation removal, stop-words removal, lemmatization and stemming [38]. Finally, in Figure 1, we present terms' coverage of papers in which bars represent

 ISBN: 978-1-61208-358-2

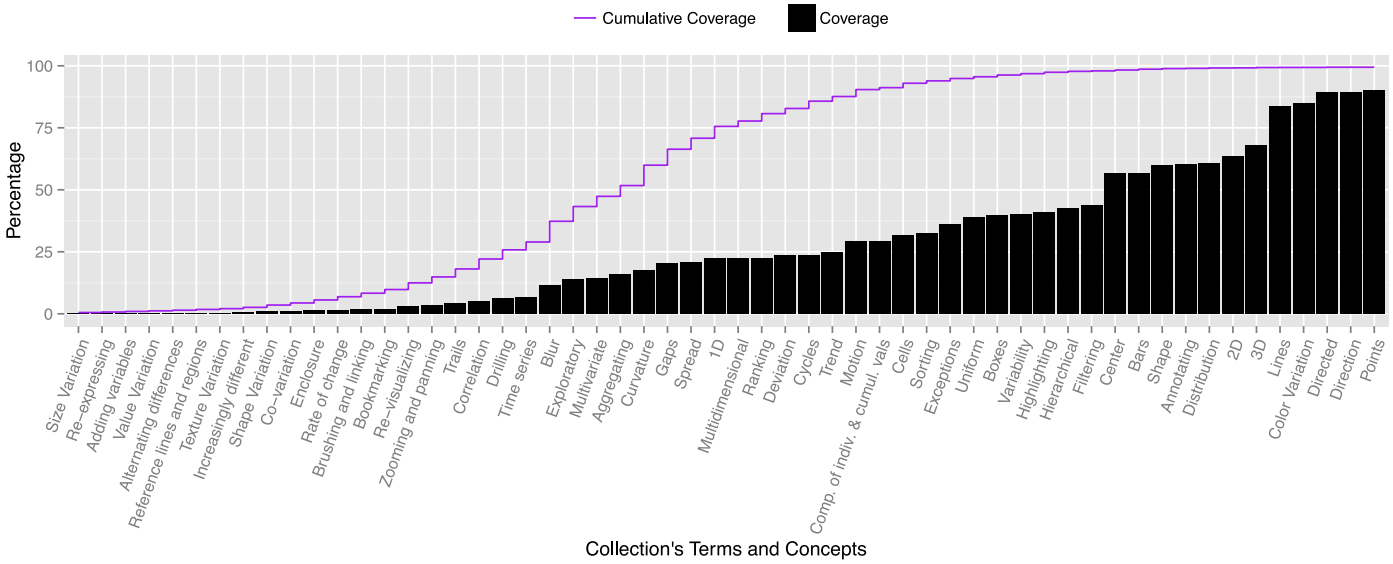


Figure 1. Terms' coverage of papers: bars represent individual terms coverage and line represents the cumulative coverage of papers (percentage) from the current term and all previous ones.

individual terms coverage (the percentage of papers in which the term appears) and line represents the cumulative coverage of papers (percentage) from the current term and all previous ones. Terms that appear in five or less papers (18) were not exposed for presentational reasons. We demonstrate that the suggested collection covered about 99% of the papers, in other words, that 99% of the papers mention at least one of the collection terms. As the most frequent terms can be very general words, we also considered the 75% of the least frequent terms. In this case, the collection still covers 94% of the papers.

V. Evaluation Results

In this section, we present some qualitative results obtained with the proposed collection of terms and concepts in the annotation of a set of visualizations as well as some quantitative results from the automatic assessment of literature coverage.

A. Use of the Annotation Process and Expressive Power of Visualization Techniques

From the 74 terms of the complete collection, 68 were used at least once to describe a visualization. The average frequency of use of a term was 27.22, the minimum was 0 and the maximum was 53, which is the number of visualizations. Hence 5 terms (*Accessing details on demand*, *Annotating*, *Bookmarking*, *Comparing* and *Filtering*) were used to describe all the techniques, which is a result of our strategy of associating each technique with every visualization capable of implementing it, even when the technique was not actually implemented. For instance, the D^3 *Line chart* has no implementation of *Details on demand*, but this analytical interaction technique could be easily implemented in that technique. There were 6 terms with no association as for instance *Texture Variation*. This lack of associations for such a small number of terms does not lower the strength of the proposed collection as the terms were all pre-attentive attributes or visual objects possibly meaning that the visualization set is not too diverse.

Figure 2 depicts the obtained annotation tree, which contains circles that represent the terms of the initial collection and squares representing each annotated visualization technique. The size and color of the squares encode, redundantly, the distance from root from dark blue (high) to light blue (low) on a continuous scale, specified by the number next to their names. Leaves that are farther from the root have more terms assigned to them and the number of terms assigned to a visualization is proportional to its *expressive power*.

On average, 27 out of 68 ($\sim 39\%$) terms are used to annotate each technique. Approximately $\sim 25\%$ of the visualizations have 19 or fewer associated terms, $\sim 50\%$ have 25 terms or fewer, 75% have 31 terms or fewer, and $\sim 90\%$ have 40 terms or fewer. Only 5 techniques are associated with more than 40 terms. We regard these 5 visualizations as special techniques concerning their high expressive power and ubiquity. These 5 techniques are all bar charts or a combination of other representations with a bar chart. The *Grouped bar chart* [39] for instance is an example of high expressive power, represented by 43 terms. At the other extreme and very close to the root of the tree, we have a *Voronoi Diagram* [40] plotted in the US map, dividing the space into a number of regions of points closer to their seed than to other seed (seeds are the US airports in 2008). Although it is a beautiful and informative visualization, it is very specific in terms of applicability.

In conclusion, in our annotation tree, a longer path between a technique and the root indicates a higher expressive power and greater potential ubiquity of that technique. The ubiquity of bar charts is well known, which in some sense demonstrates the correctness and usefulness of our methodology in analyzing this phenomenon.

 ISBN: 978-1-61208-358-2

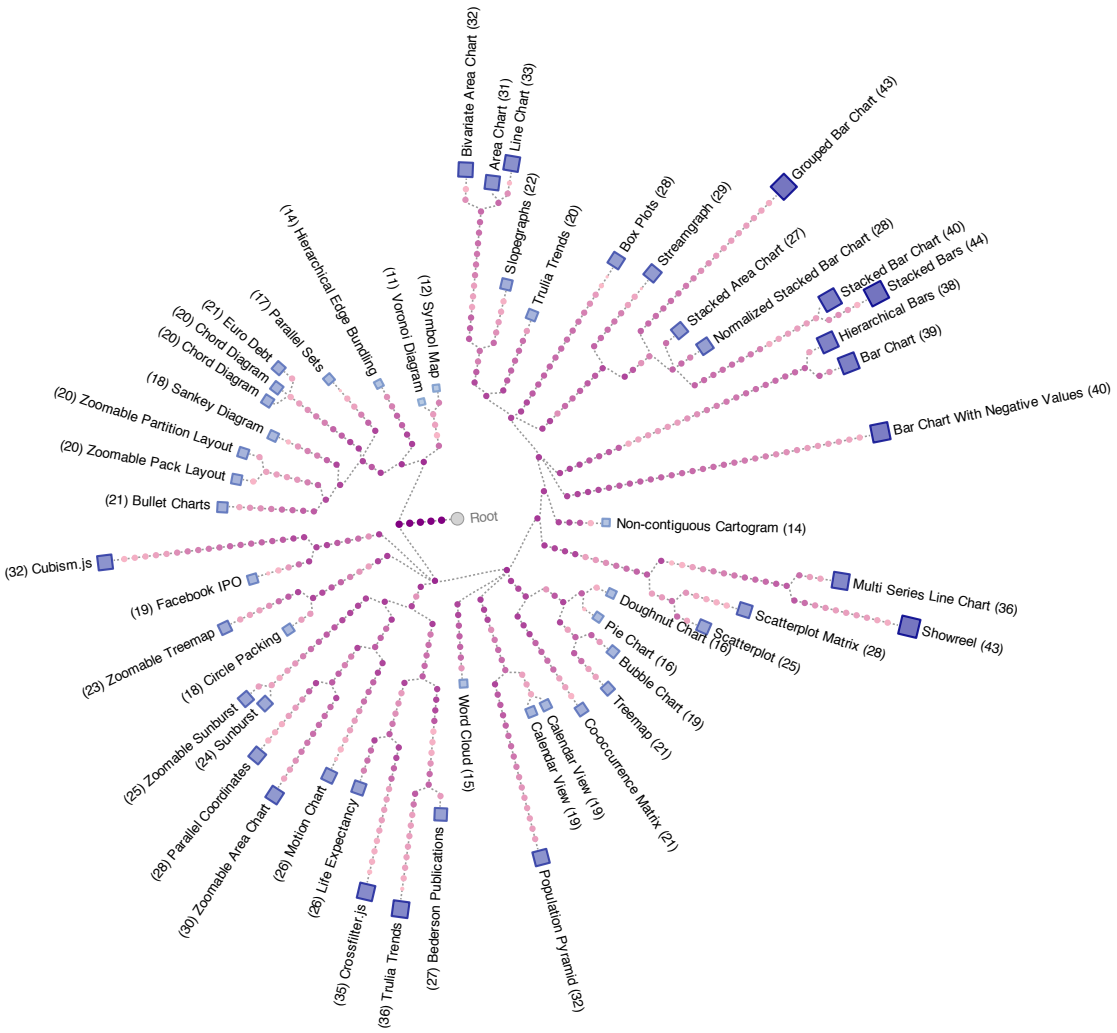


Figure 2. Modified FP-tree for annotating visualization techniques. The circles (internal nodes) represent the terms and the squares (leaves), visualization techniques. The circle colors encode the ratio *DNF* / *TNF* from purple (high) to pink (low) on a discrete scale.

B. Tree Characterization and Discriminative Power of Terms

We used the following metrics to characterize the tree and evaluate the terms of the proposed annotation process: *Dataset Node Frequency (DNF)* is the frequency of the term in the annotation of techniques in the whole dataset; *Tree Node Frequency (TNF)* is the frequency of the node representing the term in the tree, which is lower than or equal to the Dataset Node Frequency due to the compactness of FP-trees and *Mean Distance From the Root (MDR)* is the mean distance of the nodes representing the terms in the tree from the root. All the metrics have the ability to distinguish terms that are very popular in the dataset from more discriminative ones. For instance, the five top nodes of the tree (*Accessing details on demand*, *Annotation*, *Bookmarking*, *Comparing* and *Filtering*) were previously mentioned to describe every single visualization in the dataset. They are not discriminative in that they can be used everywhere and represent interesting and ubiquitous analytical interaction techniques. On the contrary, less frequent terms commonly appear far from the root and tend to be more discriminative. For instance, the term *Rate of change*, which is the percentage difference between one value to the next, presents a *MDR* of 27 and is set only for three techniques: *Line chart*, *Multi Series Line Chart* and *Showreel*, which can show the rate of change when using a logarithmic scale. The same happens for the analytical technique *Comparison of Individual and Cumulative Values* and for the visual patterns *Uniformly Different*, *Non-uniformly Different*, *Increasingly Different* and *Alternating Differences*, which we found very particular of bar charts. The term *Color Variation* is not so frequent in the dataset (60%) but is the most frequent node in the tree, appearing 18 times in various branches because it is the most used pre-attentive attribute in visualization techniques in general.

In Figure 3, we present a distribution of the values for each metric, which are all skewed. Both *DNF* and *TNF* are skewed to the left. The *DNF* has a mean of 11 and a *TNF* of 5. 95% of the terms have frequencies below 52 in the dataset, whereas 95% of the terms are presented fewer than 15 times in the tree. The compression of the tree is apparent here. *Distance from root* is skewed to the right, as the majority of the terms are far from the root, with a mean of 20.

We analyzed the tree under the perspectives of the different

 ISBN: 978-1-61208-358-2

metrics on a continuous scale ranging from higher values to lower values (results not presented due to space limitations). At a first glance, it was difficult to extract interesting patterns from the tree visualizations due to their complexity. The tree visualizations only revealed a color pattern that goes from the root to the leaves, except for some extreme cases, such as the five top nodes that have a very high frequency in the dataset.

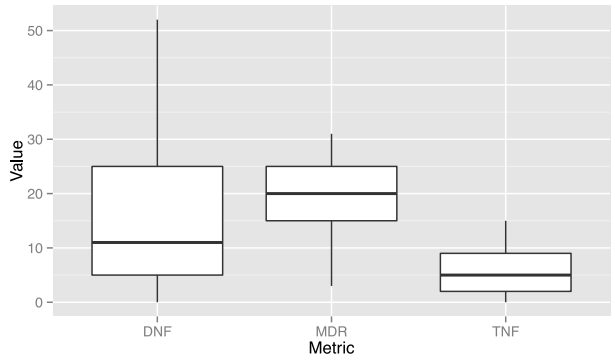


Figure 3. Distribution of the metrics: Dataset Node Frequency (*DNF*), Tree Node Frequency (*TNF*) and Mean Distance From the Root (*MDR*).

An interesting analysis came up when we colored the tree by the ratio between *DNF* and *TNF*, and the result is presented in Figure 2. When the ratio was presented on a continuous scale, its distribution was very skewed and was not easy to spot a pattern. We then used a non-uniform discretization (cuts are presented in Figure 4). The dark purple group (ratio $\geq$ 53) has already been discussed and comprises the five terms that apply to all annotated visualizations. *Exceptions*, *Directed*, *Highlighting*, *Aggregating* and *Trend* are terms presented in light purple ($4 \leq$ ratio $<$ 53), which represents highly discriminative items.

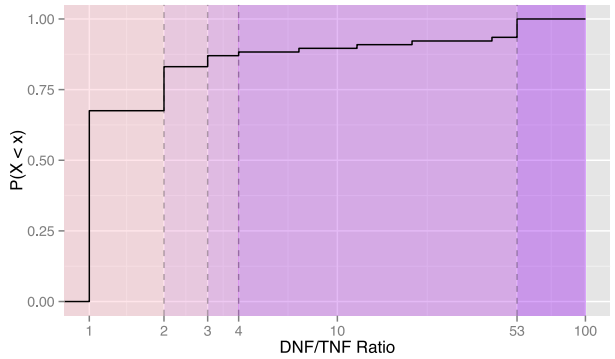


Figure 4. Discretization scheme for the *DNF* / *TNF* Ratio color scale (pink to purple). Note that x-axis is log scaled.

A broad but easily applicable characteristic of visualizations in general is how straightforward they are in communicating the underlying data and producing the desired insights; we call this characteristic *Directed* (ratio 20). The *Bubble Chart* is a good case of a visualization that does not share this characteristic, as it evolves and answers multiple questions along its dynamic life-cycle. These four attributes are comprehensive enough that they are not usually related to a single visualization, but are instead related to a large group. The darker shade of pink ($3 \leq$ ratio $<$ 4) is composed of terms that still have a large discriminative power, but already show some sort of specialization capability. *Size variation* (ratio 2.13) is a good example of this group: it is still discriminative enough to put the *Line chart* and *Bar chart* into separate groups but also specializes the whole group of bar charts (*Stacked Bar Chart*, *Hierarchical Bar Chart*), separating it from the *Streamgraph*, a "cousin" visualization that shares many terms. Terms that fall in the pink group (ratio $\leq$ 2), the largest one, do not have a strong discriminative bias to be close to the root of the tree and are sometimes very specific, being applied to a single technique. The attribute *Sorting* (ratio 1.93) is a relevant example from this group, as roughly half of the visualizations implement or could implement this functionality, but it still discriminates the *Treemap* from the *Doughnut* and *Pie Chart*. A visual object term, such as *Cells* (ratio 1.29), or a display, like *Bar graphs* (ratio 1.2), denotes high specialization.

VI. Conclusion and Future Work

We propose an embryonic version of an annotation process based on an initial collection of terms and concepts extracted from the existing literature that encompasses the visual components and capabilities of visualizations. We select a set of diverse visualizations from the D^3 gallery and annotate them with the proposed terms and concepts. We propose a visual index in form of an annotation tree that helped us to visualize the whole set of techniques and the terms associated to each of them. We characterize the proposed tree, more specifically the terms and the visualizations, concerning three metrics and were able to identify interesting patterns: the discriminative power of terms in relation to the visualizations being described and the expressive power for the visualization techniques. Qualitatively, our results demonstrate the utility of the proposed annotation process in describing visualizations as well as in understanding their capabilities and applicability. In the future, we intend to study how the proposed annotation tree can be used in automatic recommendation tasks to help users to select visualizations for specific problems and to represent data in a satisfactory way. Furthermore, to show quantitatively the completeness of the initial collection of terms and concepts, we automatically assess its coverage across all published papers from six major international information visualization conferences since 1995. Our results attest the expressiveness of the proposed collection and its coverage of over 99% of the published literature.

Finally, we acknowledge our challenge in achieving a consensus from most users of the area and our limitations concerning evaluation and evolution of the annotation process and its components. For that, we developed a platform, *CrowdVIS*, based on crowdsourcing [41]. The main goal of this platform is to use the annotation process's methodology and dynamically evolve the proposed collection of concepts and data visualization techniques, as well as their annotations [42]. Moreover, it should allow users to continuously evaluate each term and technique and to add new ones [43]. A prototype of the system is available at www.crowdvis.dcc.ufmg.br. We believe that the participation of the information visualization community, by annotating the existing visualizations in a similar way and including new visualizations in a public repository will represent a valuable contribution to future studies that could arise from ours. We intend to keep the dataset and annotations open. Certainly, this annotation process, the initial collection of terms and concepts, the annotation procedure and the dataset

 ISBN: 978-1-61208-358-2

could evolve significantly with community involvement and become intrinsic to the field in the future.

REFERENCES

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," Nature Genetics, vol. 25, no. 1, May 2000, pp. 25–29.

[2] A. Gilchrist, "Thesauri, taxonomies and ontologies _ an etymological note," Journal of Documentation, vol. 59, no. 1, 2002, pp. 7–18.

[3] B. Shneiderman, Designing the User Interface: Strategies for Effective Human-computer Interaction. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1986.

[4] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," Queue, vol. 10, no. 2, Feb. 2012, pp. 30:30–30:55.

[5] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," in Proceedings of the IEEE Symposium on Information Vizualization 2000, ser. INFOVIS '00. Washington, DC, USA: IEEE Computer Society, 2000, pp. 69–69.

[6] M. Tory and T. Möller, "Rethinking visualization: A high-level taxonomy," in Proceedings of the IEEE Symposium on Information Visualization, ser. INFOVIS '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 151–158.

[7] G. Ellis and A. Dix, "A taxonomy of clutter reduction for information visualisation," IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, Nov. 2007, pp. 1216–1223.

[8] D. Pfitzner, V. Hobbs, and D. Powers, "A unified taxonomic framework for information visualization," in Proceedings of the Asia-Pacific Symposium on Information Visualisation - Volume 24, ser. APVis '03. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2003, pp. 57–66.

[9] G. Shu, N. J. Avis, and O. F. Rana, "Bringing semantics to visualization services," Adv. Eng. Softw., vol. 39, no. 6, 2008, pp. 514–520.

[10] M. X. Zhou and S. Feiner, "Visual task characterization for automated visual discourse synthesis." in CHI, M. E. Atwood, C.-M. Karat, A. M. Lund, J. Coutaz, and J. Karat, Eds. ACM, 1998, pp. 392–399.

[11] I. Fujishiro, R. Furuhata, Y. Ichikawa, and Y. Takeshima, "Gadget/iv: A taxonomic approach to semi-automatic design of information visualization applications using modular visualization environment." in INFOVIS, J. D. Mackinlay, S. F. Roth, and D. A. Keim, Eds. IEEE Computer Society, 2000, pp. 77–83.

[12] E. Morse, M. Lewis, and K. A. Olsen, "Evaluating visualizations: Using a taxonomic guide," Int. J. Hum.-Comput. Stud., vol. 53, no. 5, Nov. 2000, pp. 637–662.

[13] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, Jan. 2004, pp. 53–87.

[14] M. Voigt and J. Polowinski, Towards a Unifying Visualization Ontology, ser. Technische Berichte. Techn.Univ., Fakultät Informatik, 2011.

[15] D. J. Duke, K. W. Brodlie, and D. A. Duce, "Building an ontology of visualization." in IEEE Visualization. IEEE Computer Society, 2004, p. 7.

[16] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in Proceedings of the 1996 IEEE Symposium on Visual Languages, ser. VL '96. Washington, DC, USA: IEEE Computer Society, 1996, pp. 336–336.

[17] S. Wehrend and C. Lewis, "A problem-oriented classification of visualization techniques," in Proceedings of the 1st Conference on Visualization '90, ser. VIS '90. Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, pp. 139–143.

[18] I. Fujishiro, Y. Takeshima, Y. Ichikawa, and K. Nakamura, "Gadget: Goal-oriented application design guidance for modular visualization environments," in Proceedings of the 8th Conference on Visualization '97, ser. VIS '97. Los Alamitos, CA, USA: IEEE Computer Society Press, 1997, pp. 245–252.

[19] O. Gilson, N. Silva, P. W. Grant, and M. Chen, "From web data to visualization via ontology mapping," in Proceedings of the 10th Joint Eurographics / IEEE - VGTC Conference on Visualization, ser. EuroVis'08. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2008, pp. 959–966.

[20] R. A. Amar, J. Eagan, and J. T. Stasko, "Low-level components of analytic activity in information visualization." in INFOVIS, J. T. Stasko and M. O. Ward, Eds. IEEE Computer Society, 2005, p. 15.

[21] E. R. Tufte, Envisioning Information. Cheshire, CT: Graphics Press, 1990.

[22] W. Cleveland, Visualizing data. AT&T Bell Laboratories, 1993.

[23] W. S. Cleveland, The elements of graphing data. Murray Hill, N.J.: AT&T Bell Laboratories, 1994.

[24] E. R. Tufte, Visual Explanations: Images and Quantities, Evidence and Narrative. Cheshire, CT: Graphics Press, 1997.

[25] S. Few, Show me the numbers : designing tables and graphs to enlighten. Oakland, Calif.: Analytics Press, 2012.

[26] E. R. Tufte, Beautiful Evidence. Graphics Press, 2006.

[27] R. Spence, Information Visualization: Design for Interaction (2Nd Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2007.

[28] C. Ware, Visual Thinking: For Design. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[29] S. Few, Now You See It: Simple Visualization Techniques for Quantitative Analysis, 1st ed. USA: Analytics Press, 2009.

[30] J. Bertin, Semiology of graphics: diagrams networks and maps. Esri Press, 2010.

[31] J. Steele and N. Iliinsky, Beautiful Visualization: Looking at Data Through the Eyes of Experts, 1st ed. O'Reilly Media, Inc., 2010.

[32] D. Wong, The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures. W W Norton & Company Incorporated, 2010.

[33] C. Ware, Information Visualization, Third Edition: Perception for Design, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.

[34] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, 2011, pp. 2301–2309.

[35] M. Bostock. Data-driven documents: Gallery. http://bit.ly/18kMazA. Accessed April, 2014. (2012)

[36] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 1–12.

[37] D. E. V. Pires, L. C. Totti, R. E. A. Moreira, E. C. Fazzion, O. L. H. M. Fonseca, W. M. Jr., R. C. de Melo Minardi, and D. O. G. Neto, "Fpcluster: An efficient out-of-core clustering strategy without a similarity metric," JIDM, vol. 3, no. 2, 2012, pp. 132–141.

[38] R. A. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[39] M. Bostock. Data-driven documents: Grouped bar chart. http://bit.ly/QsqidV. Accessed April, 2014. (2012)

[40] M. Bostock. Data-driven documents: U.s. airports, 2008 voronoi diagram. http://bit.ly/1ttgoI4. Accessed April, 2014. (2012)

[41] J. Howe, "The rise of crowdsourcing," Wired magazine, vol. 14, no. 6, 2006, pp. 1–4.

[42] D. Karampinas and P. Triantafillou, "Crowdsourcing taxonomies," in The Semantic Web: Research and Applications, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7295, pp. 545–559.

[43] J. Mortensen, "Crowdsourcing ontology verification," in The Semantic Web – ISWC 2013, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 8219, pp. 448–455.

 ISBN: 978-1-61208-358-2

Prediction Model Framework for Imbalanced Datasets

Maria Rossana C. de Leon
Southern Luzon State University
Lucban, Quezon Philippines
e-mail: rossana.4481@gmail.com

Eugene Rex L. Jalao
University of the Philippines Diliman
Quezon City Philippines
e-mail: eljalao@upd.edu.ph

***Abstract*—Generally, prediction requires significant and good quality input data that will give accurate prediction. However, real-data are often noisy, inconsistent, and imbalanced. If the classes are imbalanced, the class accuracy is unlikeable because the prediction tends to favor those in majority class since it has relatively significant class size. To resolve the imbalance problem, a resampling algorithm is proposed which improves the prediction accuracy of each class. The algorithm was tested in 4 different datasets each using different prediction and classification methodologies, such as Regression Analysis, Decision Tree, Rule Induction, and Artificial Neural Networks. Results show that the framework works in either methodologies and prediction accuracy generally improves after resampling. The framework was also compared to the existing sampling methodologies and results show that it is comparable with the ROS/RUS, but the resampling rate is minimized with the proposed framework.**

Keywords:Prediction model framework; Class imbalance problem; Data mining.

I. INTRODUCTION

Prediction, pattern recognition, and classification problems are not new. By definition, predictive analytics is the utilization of statistics, data mining, and game theory to analyze current and historical facts in order to make predictions about future events [1]. It enables decision makers to develop mathematical models to help them better understand the relationship among variables.

One of the major issues in coming up with accurate predictions lies in the quality of input data, which are usually *incomplete* (lacking attribute values or certain attributes of interest, or containing only aggregate data), *noisy* (containing errors, or *outlier* values that deviate from the expected), *inconsistent* (e.g., containing discrepancies in the department codes used to categorize items), and *imbalanced* (occurs when one class is underrepresented in the data set). Accordingly, low quality data will lead to low quality prediction and classification results [2]. This research mainly focuses on addressing imbalanced datasets.

Generally, a two-class data set is said to be imbalanced (or skewed) when one of the classes, called the *minority class*, is heavily under-represented in comparison to the other class, called the *majority class*. Dataset imbalance on the order of 100 to 1 is prevalent in fraud detection and imbalance of up to 100,000 to 1 has been reported in other applications [3]. In such a situation, most of the classifiers are biased towards the major classes and hence show very poor classification rates on minor classes. It is also possible that a classifier predicts everything as major class and ignores the minor class [4]. Class distribution, i.e., the proportion of instances belonging to each class in a data set, plays a key role in classification. Data sets with skewed class distribution usually tend to suffer from class overlapping, small sample size or small disjuncts, which difficult classifier learning [5]. Furthermore, the evaluation criterion can lead to ignore minority class examples (treating them as noise) and hence, the induced classifier might lose its classification ability in this scenario. In many applications, misclassifying a rare event can result in more serious problem than common event. For example, in medical diagnosis in case of cancerous cell detection, misclassifying non-cancerous cells leads to some additional clinical testing but misclassifying cancerous cells leads to very serious health risks. However, in classification problems with imbalanced data, the minority class examples are more likely to be misclassified than the majority class examples, due to their design principles; most of the machines learning algorithms optimize the overall classification accuracy which results in misclassification minority classes.

Various studies have already been conducted that answers class imbalance problem. Yet, the techniques are either simple or complex. Simple techniques use either oversampling or undersampling or combination of both, but the techniques usually assume that a fully balanced dataset can be attained. With this assumption, the chance of over-fitting, particularly for oversampling, is highly possible. In the case of undersampling, too much useful data that are excluded in the training set can make the prediction or classification inaccurate. Multi-dimensional (*n*-dimension) data sets can be resolved using more sophisticated techniques which can be hard to interpret and would require a significant amount of processing cost.

Against this background, there is a need to develop a prediction model framework that can pre-process and resolve the imbalance problem by utilizing a proposed iterative oversampling and undersampling methodology for *n*-dimensional datasets.

We begin by presenting related works of others in Section 2. In Section 3, we describe the proposed framework and the algorithm to resolve the class imbalance

 ISBN: 978-1-61208-358-2

problem. Then, the results of the experimental runs and the analyses are discussed in Section 4. Finally, conclusion and areas for future studies are described in Section 5.

II. RELATED WORKS

Large numbers of approaches have previously been proposed to deal with the class-imbalance problem [6]. The approaches are categorized into two groups: the *internal approaches acting on the algorithm* that create new algorithms or modify existing ones to take the class-imbalance problem into consideration, and *external approaches acting on the data* that use unmodified existing algorithms, but resample the data presented to these algorithms so as diminish the effect caused by their class imbalance [1]. Internal approaches modify the learning algorithm to deal with the imbalance problem. They can adapt the decision threshold to create a bias toward the minority class or introduce costs in the learning process to compensate the minority class. Cost sensitive learning is probably the most well-known method of dealing with the class imbalance [3]. External approaches act on the data instead of the learning method. They have the advantage of being independent from the classifier used.

Sampling is the most popular means for overcoming the class imbalance problem [3]. Sampling is used as a means of altering the distribution of the minority class so that it is not under represented when training the learner. There are three basic approaches to overcome the class imbalance problem. These are over sampling of the minority class, under sampling of the majority class or the use of a hybrid approach based on both.

A. Random Over-sampling (ROS)

ROS [3] can be described as the random sampling of the minority class with replacement. This randomly samples with replacement the minority class and adds them to the minority class sample set until the size of the minority class is the same size as the majority class. With replacement means that after each resample, the samples are placed back in the 'pot' (the minority sample set) and can be resampled again. Over-sampling can result in a number of problems, these include over-fitting of a model especially in the cases of noisy data. Also, over-sampling does not result in more information being included in the training set, which can cause the production of overly complex models.

B. Random Under-sampling (RUS)

When under sampling [11] is used to overcome the problem of class imbalance, the number of majority class examples is reduced until the number of majority samples equals the number of minority samples. When using this solution to the class imbalance problem, certain problems may arise from removing such a larger number of the majority class, due to a fact that a number of potentially useful samples from the majority class may be discarded. Under-sampling does offer a number of benefits to over-sampling. The main one being that it results in a smaller training set as compared to oversampling, thus resulting in shorter training times.

C. Combination of ROS and RUS (ROS/RUS)

ROS/RUS is a combination of random over-sampling and random under-sampling. When this approach is used, the majority class would be under-sampled and the minority class would be over-sampled.

D. Synthetic Minority Over-sampling Technique (SMOTE)

The use of SMOTE algorithm [9] to artificially synthesize items belonging to the minority class has also been postulated as a means of overcoming the class imbalance problem. When SMOTING a dataset, the class having the smaller number of examples is over-sampled through the synthesis of artificial instances, as opposed to over-sampling existing samples with replacement. The class having the smaller number of examples is over-sampled by the use of a *kNN* to add artificial instances along the line segments connecting some or the entire population of nearest neighbors. The number of nearest neighbors to add is dependent on the level of over-sampling required.

E. Cost-sensitive Learning

Cost-sensitive learning framework incorporates both data level transformations (by adding costs to instances) and algorithm level modifications (by modifying the learning process to accept costs) [1]. It biases the classifier toward the minority class the assumption higher misclassification costs for this class and seeking to minimize the total cost errors of both classes. The major drawback of these approaches is the need to define misclassification costs, which are not usually available in the data sets [5].

Several studies that used the abovementioned approaches have been done to answer the class imbalance problem. Some of which are discussed below:

Kubat and Matwin [4] selectively under-sampled the majority class while keeping the original population of the minority class. The minority examples were divided into four categories: some noise overlapping the positive class decision region, borderline samples, redundant samples and safe samples. The borderline examples were detected using the Tomek links concept.

Japkowicz [8] discussed the effect of imbalance in a dataset. She evaluated three strategies: under-sampling, resampling and a recognition-based induction scheme. She experimented on artificial 1D data in order to easily measure and construct concept complexity. Two resampling methods were considered. Random resampling consisted of resampling the smaller class at random until it consisted of as many samples as the majority class and focused resampling consisted of resampling only those minority examples that occurred on the boundary between the minority and majority classes. Random under-sampling was considered, which involved under-sampling the majority class samples at random until their numbers matched the

 ISBN: 978-1-61208-358-2

number of minority class samples; focused under-sampling involved under-sampling the majority class samples lying further away. She noted that both the sampling approaches were effective, and she also observed that using the sophisticated sampling techniques *did* not give any clear advantage in the domain considered.

Another study used the cost-sensitive learning, as cited by Chawla [9]. He compares the "meta-cost" approach to each of majority under-sampling and minority over-sampling. He finds that meta-cost improves over either, and that under-sampling is preferable to minority over-sampling. Error-based classifiers are made cost-sensitive. The probability of each class for each example is estimated, and the examples are relabeled optimally with respect to the misclassification costs. The relabeling of the examples expands the decision space as it creates new samples from which the classifier may learn [3].

Estrabooks [6] used the external approach to resolve the class imbalance problem. Resampling was conducted using the following strategies: over-sampling consisted of copying existing training examples at random and adding them to the training set until a full balance was reached. Under-sampling consisted of removing existing examples at random until a full balance was reached. The results suggested the neither over-sampling nor the under-sampling strategy is always the best one to use, and finding a way to combine them could perhaps be useful, especially if the bias resulting from each strategy is of a different nature [4]. Furthermore, the study suggested that resampling to full balance is generally not the optimal resampling rate, at least when the test set is balanced. The optimal resampling rate varies from domain to domain and resampling strategy to resampling strategy. In general, over-sampling changes its effect gradually and in a stable manner with different rates, while under-sampling does so radically and in an unstable manner.

Brennan [3] made a survey on the methods for overcoming the class imbalance problem in fraud detection. RapidMiner, R and Weka were used to study the various methods for overcoming the class imbalance problem. All the sampling methodologies and the cost-sensitive learning method were applied in three different datasets such as car insurance fraud dataset, consumer fraud insurance dataset, and thyroid disease dataset. For all datasets used, the data methods proved to be superior to the algorithmic methods. The data methods surveyed were found to be simple to implement and at least some of them were highly effective. They also proved easier to implement and did not lead to sizable increases in training time or resources needed. SMOTE, ROS, and ROS/RUS proved to be the best performing methods for treating the imbalance in the data. However, ROS may be criticized as a method as it can lead to over-fitting a model as it over trains a model to recognize a small number of minority classes. SMOTE addresses this by creating artificial replicants and thereby creating a less specific feature space for the minority group. However, if the SMOTE algorithm has proved ineffectual at replicating the characteristics of the minority class, it will result in a situation where the minority class sample qualities are too similar to those of the majority class. This result in the problem of "class mixture" and the resulting model will misclassify classes of the minority class as members of the majority class.

This study focuses on the data methodologies (external approach) because of the advantages that it offers and literature argued that this approach is better than the algorithmic methodologies (internal approach). However, the existing data methodologies only assume a random sampling technique that *fully* balances the number of examples in the majority and minority classes.

Furthermore, literature suggests that external approach may be divided into two types of categories. First, there are approaches that focus on studying what the best *data* for inclusion in the training set [10], and second, there are approaches that focus on studying what the best *proportion* of positive and negative examples to include in training set [11]. The work of Estrabooks [6] focused on the second category by creating a framework that deals with the proportion question. This study attempts to answer both questions. The suggested framework provides the best *rate* at which with the combined over-sampling and under-sampling methodology will provide good prediction accuracy even without fully balancing the number of examples in the minority and majority classes. At the same time, it will provide a methodology in determining the data for inclusion in the training set for the over-sampling and the data for exclusion in the training set for the under-sampling methodology.

III. PROPOSED ALGORITHM

This research proposes a combination of oversampling and undersampling methodologies to determine the appropriate number of samples in each class that will give higher class accuracy.

Suppose we have a dataset represented by matrix $\boldsymbol{A}$ with a set of row r and column c. The rows represent examples and the columns are the attributes of the examples with d dimensions. Thus, the matrix element r_{rc} is the value of example with ID r in the attribute with ID c. Consider such a matrix $\boldsymbol{A}$, with n rows and m columns, defined by its set of rows, $R = \{r_1, \ldots, r_n\}$, and its set of columns, $C = \{c_1, \ldots, c_m\}$. Thus, matrix $\boldsymbol{A}$ can be denoted by (R, C). This study provides a framework that can predict variable y, which is directly or indirectly affected by the attributes defined by $C = \{ c_1, \ldots, c_m\}$. If after discretizing the predictor variable y, matrix $\boldsymbol{A}$ can be partitioned into k classes with n_1 elements in class 1… n_k elements in class k, where $n_1, n_2, \ldots, n_k$ are not uniformly distributed. We can define a majority class if $n_k = n_{\text{maximum}}$, otherwise n_k is called minority class. Discussed below is the proposed resampling algorithm to resolve the class imbalance problem.

 ISBN: 978-1-61208-358-2

Algorithm
If matrix ***A***, the original dataset, is divided into *k* classes: $\{C_1, C_2, \ldots, C_k; n_1, n_2, \ldots, n_k\}$
S% = resampling rate
m = number of minority class samples
M = number of majority class samples
n_t = number of examples in class *k* at *i*th iteration.
n_{t-1} = number of examples in class *k* at previous iteration

1 Initialize *S*% = 0.
2 Choose class C_k, where $C_k \subseteq \boldsymbol{A}$
3 If C_k is minority class, then oversample the C_k class.
4 For the array of original minority class samples of size *m*, assign a two-digit number, initialized to 01.
5 Generate random numbers, between 0 to 1, equal to the required oversample size.
6 Use the first two digits to create the synthetic sample of size n_k from the generated random numbers.
7 If C_k is majority class, then undersample the C_k class.
8 Compute $i = \frac{n_{t-1}}{(n_{t-1}-n_t)}$.
9 For the array of original majority class samples of size *M*, eliminate every *i*th instance from the data set.
10
11 Predict *y*
12 Compute over-all accuracy and class accuracy
13 While class accuracy improves
14 Increment *S*%
15 Go to 3
16
17 Stop

Resampling has been conducted using the following strategies: oversampling consisted of copying existing training examples at random and adding them to the training set; while undersampling consisted of removing existing examples at random until the desired number of samples is reached. For example, if datasets initially has 1000 samples in majority class and 50 samples in minority class, that is at *S* = 0%. Using a 10% resampling rate will contain 1000 – [0.10*(1000-50)] = 905 majority class examples and $50 + (0.10 * 1000) = 150$ samples in the minority class. Lines 4 – 6 of the abovementioned algorithm show how oversampling is being done. Oversampling starts by assigning a two-digit number to each of the original minority examples, since the number of minority class is usually less than 100, for most situations. However, the user can redefine it depending on the number of minority examples. By generating random numbers, which is equal to the desired number of examples defined by *S%*, copy the examples at random and add them to the training set. Conversely, lines 8 – 9 show the undersampling procedure. Computing the *i* determines the dataset that will be excluded from the original majority class. The datasets will therefore be resampled by simultaneously reducing the number of majority examples and increasing the number of minority examples. Thus, the algorithm will make the dataset approximately uniformly distributed without abruptly changing the class size, hence, overfitting can be avoided, as well as removing the important data in the majority class can also be prevented. The algorithm terminates if the class accuracy starts to plateau, meaning, the accuracy stops improving. Iterating it further will not improve the accuracy. It therefore implies that the iteration stops at the same time, that is, to minimize the computational time of the algorithm. By applying a combination of undersampling and oversampling, the initial bias of the learner towards the majority class is avoided.

This study assumes that the framework can be applied to any classification/prediction methodologies. Some methodologies that have been investigated are regression, decision tree induction, rule induction, and artificial neural network.

The performance of the prediction/classification is evaluated by a confusion matrix [3], as illustrated in Figure 1.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 1. Confusion Matrix

The confusion matrix is a useful tool for analyzing how well the classifier can recognize examples of different classes. The columns are the Predicted class and the rows are the Actual class. In the confusion matrix, TN is the number of negative examples correctly classified (True Negatives), FP is the number of negative examples incorrectly classified as positive (False Positives), FN is the number of positive examples incorrectly classified as negative (False Negatives) and TP is the number of positive examples correctly classified (True Positives). Predictive accuracy is the performance measure generally associated prediction or classification algorithms and is defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

In the context of balanced datasets and equal error costs, it is reasonable to use error rate as a performance metric. Error rate is 1 – *Accuracy*.

IV. RESULTS AND DISCUSSION

A. Test Case Data

The proposed algorithm has been tested on four test data sets with various features. The first two data sets are original datasets while the last two data sets are from Knowledge Extraction based on Evolutionary Learning

(KEEL) Data Repository [14]. Table I lists down the four data sets and their corresponding properties, while Figure 2 shows the distribution histogram of the classes of each dataset.

TABLE I – COMPARISON OF TEST CASE DATA

Data Set	Number of Instances	Number of Predictor Variables	Number of Classes	Proportion of Class Imbalance
Crop Yield	1,003	21	3	10:1:1
Credit Risk	348	19	2	3:1
Ecoli	336	7	2	16:1
Yeast	1,489	8	10	90:50:10:1

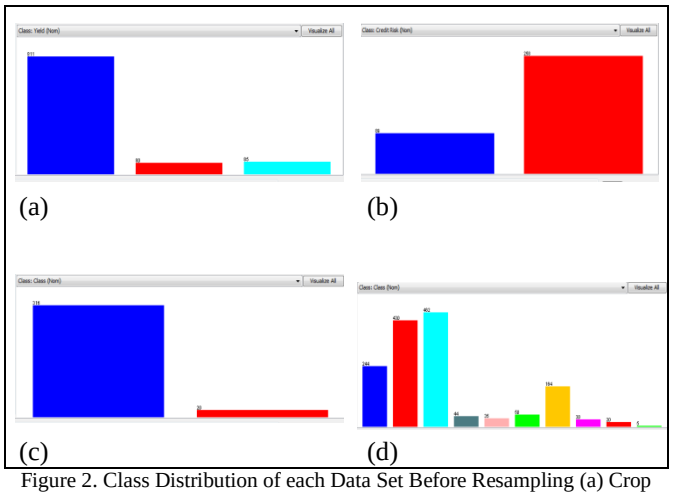


Figure 2. Class Distribution of each Data Set Before Resampling (a) Crop Yield, (b) Credit Risk, (c) Ecoli, and (d) Yeast

B. Prediction

As mentioned in the previous section, various prediction/classification methodologies have been applied, such as regression analysis, decision tree, rule induction technique, and artificial neural network. Data mining algorithms are applied using Waikato Environment for Knowledge Analysis (WEKA) version 3.6.10 [14] software for decision tree, rule induction, and neural network. It includes a wide variety of learning algorithms and preprocessing tools. Minitab 16 has been used to implement the regression analysis.

The performance of the applied data mining algorithms is estimated by the 10-fold cross validation. Data are randomly partitioned into 10 blocks, one block is held out for the test purpose and the model is built on the remaining nine blocks. This method is then repeated for other blocks. After repeating the calibration and validation processes with ten different combinations, the results (prediction accuracy and MAE) obtained with these ten different validation datasets are summarized by calculating the mean value and 95% confidence interval. It should be noted, however, that the calibration and validation dataset are independents throughout the procedure.

C. Resampling Rate

Resampling rate is a parameter that will be selected by the user (user-specified). Choosing the appropriate resampling rate and its increment size define the computational time in achieving the balanced dataset. Selecting a low increment size might mean a slow convergence while incrementing it at high values might show a very fast convergence, which in effect might not give the best resampling size. Figure 2 shows the summary of the analysis.

It is gleaned from Fig. 3 that incrementing it at 1% shows a very slow convergence, which is after 29 runs. However, having a step increment of 20% shows a very fast convergence, which only took 3 runs. This is not good because it will not give us the appropriate resampling size.

A resampling size is said to be appropriate if it is not exceedingly under-sampled nor oversampled. Generally, increment size of the resampling rate depends on how large or small is the gap of the majority class and the minority class. The smaller the gap between the majority and minority class means small increment size for resampling rate can be used. However, if the disparity is high, higher size increment is suggested.

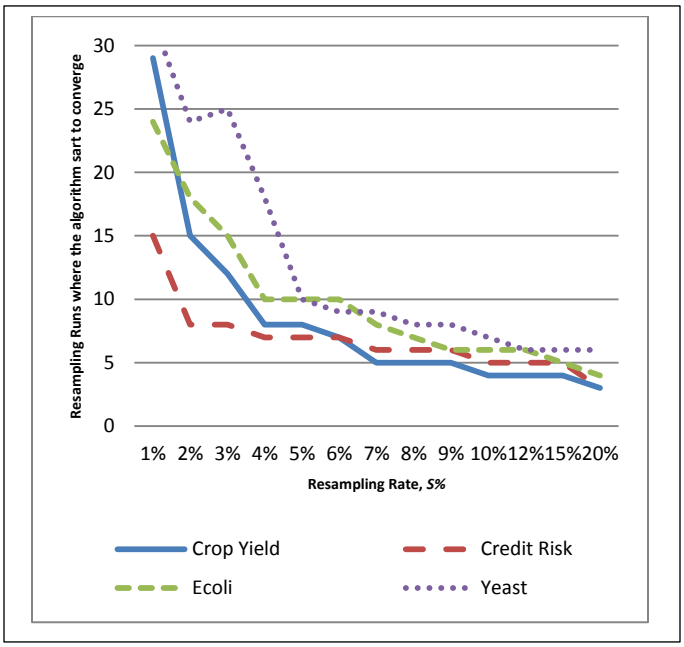


Figure 3. Convergence at Different Resampling Rate

D. Resampling Size

After identifying the increment size of the resampling rate, we can now determine the appropriate resampling size. As mentioned above, a good resampling size will give the user an approximately balanced class size that is not overly undersampled nor oversampled, thus, overfitting can be precluded. Table II shows the different results of the resampling runs.

 ISBN: 978-1-61208-358-2

TABLE II – RESAMPLING RUNS

Resampling Size	Data Sets			
	Crop Yield	*Credit Risk*	*Ecoli*	*Yeast*
Class A	568	232	269	370
Class B	328	115	67	344
Class C	323	-	-	288
Class D	-	-	-	224
Class E	-	-	-	132
Class F	-	-	-	128
Class G	-	-	-	116
Class H	-	-	-	120
Class I	-	-	-	108
Class J	-	-	-	96
Ressampling Rate	5%	2%	5%	5%
Number of Resampling Runs to Converge	8	4	7	6

It is shown in Table II that the resampling rate of 5% implies that each run increases the undersampling or oversampling rate by 5%. Most of the data sets in this research use a resampling rate of 5%. The number of resampling runs shows the number of runs by which the accuracy graph converges. After that run, the next runs show very little improvement in each class accuracy, that is the graph starts to plateau. Hence, it is considered as the appropriate resampling size for each data set. Fig. 4 shows the improved class distribution after resampling.

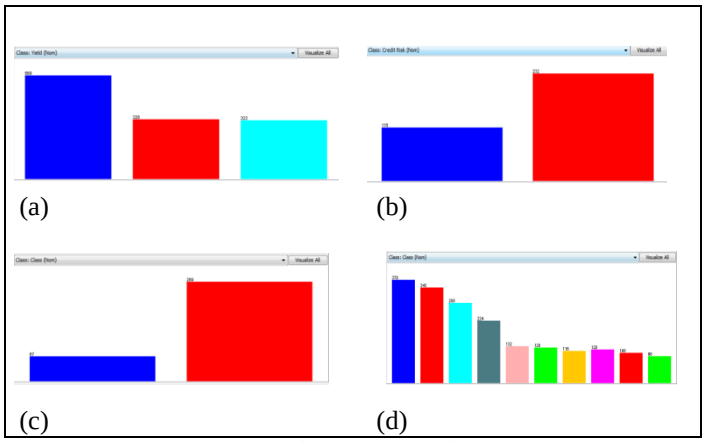


Figure 4. Class Distribution of each Data Set After Resampling (a) Crop Yield, (b) Credit Risk, (c) Ecoli, and (d) Yeast

The class accuracy generally improves for each data set after resampling, using any of the prediction/classification methodologies. Figure 5a to Figure 5d show the comparison of the overall accuracy and class accuracy before and after the resampling using the regression analysis, decision tree induction, rule induction, and artificial neural network.

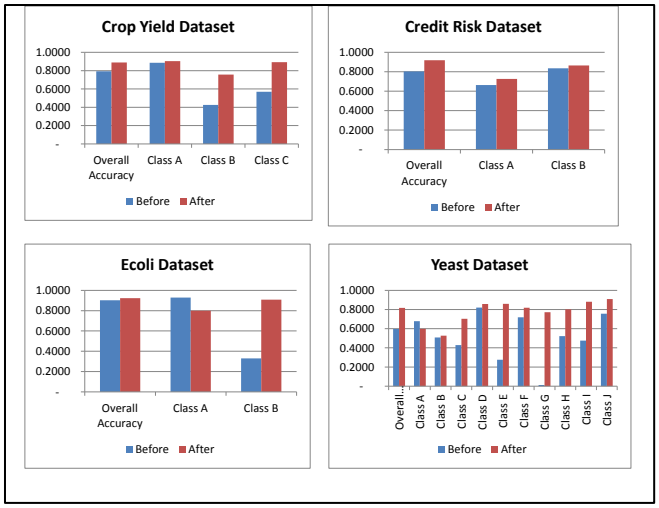


Figure 5a. Comparison of Over-all and Class Accuracy Before and After Resampling Using Regression Analysis

Regression analysis is implemented using Minitab 16 software. Since the predictor variable is discretized, general regression is used. A regression model is developed initially using the original dataset (before resampling) for each dataset. For the Crop Yield dataset, although the overall accuracy is almost 80%, investigating the class accuracy shows that Class B is the problematic class since only 42% was correctly classified examples. Class C is also problematic since it only 57% was correctly classified. The low prediction accuracy for these classes is caused by the fact that Classes B and C are both minority class. Credit Risk dataset has little problem on class A accuracy since the accuracy is almost 66% as compared to class B accuracy which is 84%. Notice that Class A is the minority class in this case, hence low accuracy is explained by this fact. For the Ecoli dataset, Class B has the lowest accuracy, which is only 33% as compared to Class A accuracy of 93%. It is because Class B is the minority class. For Yeast dataset, five classes have class accuracy of below 50%, which are Classes C, E, G, H, and I, Class G being the most problematic with class accuracy of only 1.19%.

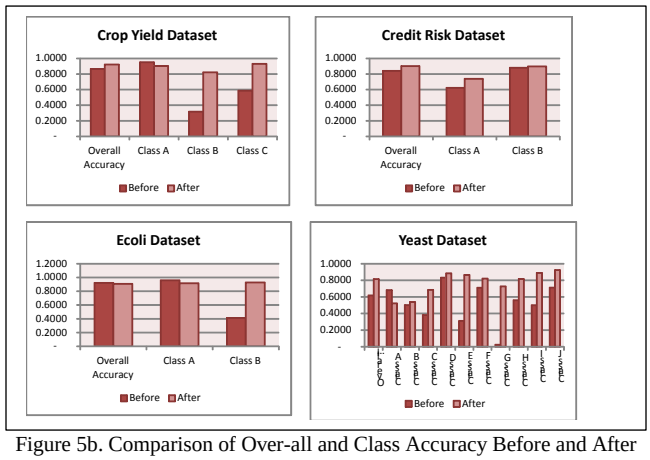


Figure 5b. Comparison of Over-all and Class Accuracy Before and After Resampling Using Decision Tree Induction Technique

The datasets are also analyzed using decision tree technique. The situation before resampling is almost the same as that of the previous technique, except that the prediction accuracy is relatively higher in using this technique. However, the minority classes are still the classes with low class accuracy. To resolve the imbalance, the proposed framework is then applied, this time using decision tree technique. Decision tree technique is applied using C4.5 algorithm, which is a built-in algorithm in WEKA. As can be seen in Figure 5b, class accuracy is again improved significantly.

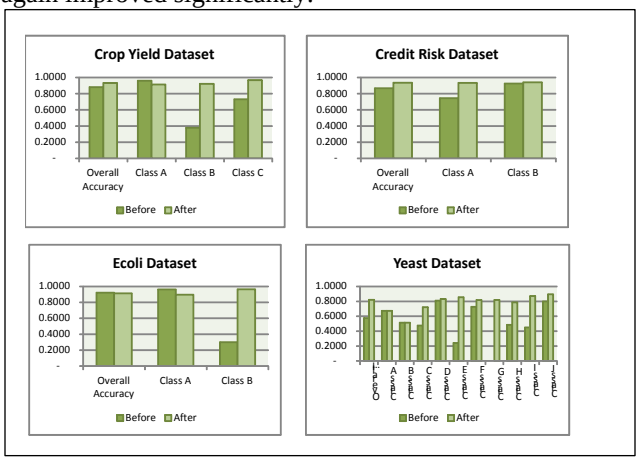


Figure 5c. Comparison of Over-all and Class Accuracy Before and After Resampling Using Rule Induction Technique

Rule Induction technique is also employed to the 4 datasets to determine if the framework can also be applied. JRip algorithm is used under this technique, which is also implemented using WEKA. Similarly, the situation is also the same with respect to class accuracy. Minority classes have the low accuracy. Resampling the class size again improves the class accuracy as shown in Figure 5c.

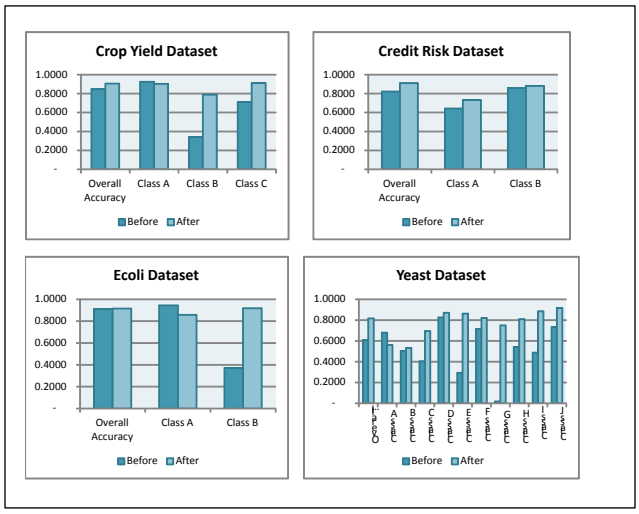


Figure 5d. Comparison of Over-all and Class Accuracy Before and After Resampling Using Artificial Neural Network

Another commonly used technique in classification is the Artificial Neural Network (ANN) [2]. Backpropagation algorithm is used under this technique. Same scenario can be observed even this technique is used, the minority classes are the classes that have the low accuracy. The framework is again applied, this time using the ANN. As can be seen in Figure 5d, the class accuracy also improved by resampling the class size.

E. Validation

The performance of the proposed framework is compared to the existing sampling methodologies used in resolving the class imbalance problem. As discussed in the literature, ROS, RUS, ROS/RUS, and the SMOTE are the most commonly used under resampling techniques. ROS has been implemented by randomly creating a number of replicants of the minority class without replacement until it is equal to the number of samples of the majority class. RUS has been implemented by simply choosing a random sample of the majority class which matches the number of minority class examples. The ROS/RUS involves the combination of the two prior methods until the number of minority samples equals the number of majority samples in the ratio 50/50. SMOTE has been implemented through WEKA, where it consists of reiterating the framework until the prediction accuracy stops improving.

The proposed framework is also compared to one of the most commonly used algorithmic technique, which is the cost-sensitive learning (CSL). Cost-sensitive learning has been implemented using Cost-Sensitive Rough Sets (COSER) in WEKA.

The methodologies are compared based on the oversampling rate, undersampling rate, elapsed time, and prediction accuracy. Oversampling rate is the percentage by which the minority class is oversampled to be equal to that of the majority class, while the undersampling rate is the percentage by which the majority class is undersampled to be equal to the minority class. Certainly, we want to minimize both the undersampling rate and the oversampling rate in the shortest possible time, so that both the overfitting and loss of data issues are avoided. Elapsed time refers to the total time by which the methodology used is able to reach the predicted accuracy. Times are measured in minutes. The prediction accuracy is the performance measure used in comparing the methodologies. The results are summarized in Table III.

Examination of the table shows that ROS has the best prediction accuracy. Similarly, ROS has the shortest elapsed time among all the dataset tested. The success of the ROS method can be explained by the fact that all the members of the majority class are being utilized and the random replication of the minority class until it is balanced with the majority class. This redistribution of the class size provides the prediction/classification learner sufficient samples to be able to train a model to recognize the minority class, and not treating them as noise. However, the near perfect performance of the ROS method must be tempered by the fact that oversampling can result in overfitting [3].

 ISBN: 978-1-61208-358-2

TABLE IV – COMPARISON WITH THE EXISTING METHODOLOGIES

		Ratio of Major class and Minor class(Cost Ratio)	Largest Down sampling	Largest Upsampling Rate	Elapsed Time (minutes)	Predicted Accuracy
Crop Yield	ROS	1.00	-	1013.75%	2.54	0.9716
	RUS	1.00	90.14%	-	4.23	0.7561
	Hybrid	1.00	49.94%	199.75%	4.89	0.9262
	SMOTE	1.19	-	750.00%	16.71	0.8315
	CSL	0.40	-	-	40.25	0.8815
	Proposed	1.76	29.96%	247.26%	24.35	0.9310
Credit Risk	ROS	1.00	-	289.88%	0.72	0.9893
	RUS	1.00	65.50%	-	1.46	0.6157
	Hybrid	1.00	50.00%	144.94%	1.18	0.9012
	SMOTE	2.35	-	23.60%	0.53	0.7647
	CSL	0.125	-	-	26.87	0.6667
	Proposed	2.02	10.08%	224.35%	2.74	0.9345
Ecoli	ROS	1.00	-	1580.00%	1.63	0.9914
	RUS	1.00	93.67%	-	1.84	0.2354
	Hybrid	1.00	50.00%	790.00%	2.26	0.8831
	SMOTE	7.90	-	100.00%	0.90	0.8324
	CSL	0.05	-	-	18.29	0.6502
	Proposed	4.01	14.87%	335.00%	3.82	0.9137
Yeast	ROS	1.00	-	9240.00%	20.78	0.9865
	RUS	1.00	89.18%	-	31.52	0.5327
	Hybrid	1.00	50.00%	462.00%	25.66	0.8329
	SMOTE	4.96	-	1820.00%	42.69	0.5634
	CSL	0.005	-	-	153.74	0.3160
	Proposed	3.85	19.91%	192.00%	45.21	0.8208

RUS delivered the poorest prediction accuracy among the dataset tested. The effect of reducing the number of majority samples results in the learner not having enough majority samples to train an effective model. This is due to the learner not being exposed to enough of the majority samples in the training set and ignoring most of the population of the majority class examples, as they are simply discarded, while training the model [3].

The ROS/RUS tends to have good performance in terms of prediction accuracy and elapsed time. It is because, to a less extent, only some of the majority examples are being removed, causing the synthesis of a less than perfect model. This technique is also better than ROS since it will avoid the chance of overfitting since the minority class will not be overly sampled.

SMOTE is also ineffectual at creating artificial replicants of the minority class. This is due to the artificial replicants it created based on the minority class are too similar to the majority class [2].

The cost-sensitive learning has not proved to be capable of improving the prediction accuracy across all the datasets used. They tend to be poor at differentiating the majority from the minority class and misclassifying the majority class as minority class members (false positives). This led to poor decision of the model. It has also been difficult to implement than the resampling methods. As for each learner, the cost ratio of the majority to the minority class misclassification cost, had to be derived empirically. It explains the long elapsed time for cost-sensitive learning methodology. Furthermore, time complexity tends to increase as the number of classes and the proportion of class imbalance increase.

The proposed framework is almost comparable to that of the hybrid technique. It can predict with a higher degree of accuracy and at the same time avoids the possibility of overfitting, as in the case of ROS. It is a better technique since it aims to determine the best class size that will give a higher prediction accuracy even without overly oversampling nor overly undersampling the classes. However, one limitation of the framework is that it takes longer time in realizing good prediction accuracy. It is because the algorithm takes more iterations than the simple ROS, RUS, and ROS/RUS since it does not require the classes to be fully balanced. Hence, it tries to determine the appropriate size for each class so that overfitting can be avoided. Furthermore, it tries to determine the appropriate sample to be included in the resampled class. In general, the proposed prediction framework for imbalanced dataset shows satisfactory performance as compared to the existing methodologies.

V. CONCLUSION AND AREAS FOR FURTHER STUDY

Literature proved that sampling methodologies are better than the algorithmic methodologies in resolving the class imbalance problem. Among the sampling methodologies being used are Random Oversampling (ROS), Random Undersampling (RUS) Hybrid technique (ROS/RUS), and the SMOTE. Nevertheless, the existing approaches for sampling have disadvantages. The questions on how much should we oversample and undersample, and which data must be included/excluded in the dataset still exist. Hence, the proposed prediction framework is developed to provide answers to these issues. This study generally aims to develop a prediction model framework that can pre-process data and resolve the imbalance problem by utilizing a proposed iterative oversampling and underampling methodology for *n*-dimensional data sets.

The framework consists of pre-processing component, which consists of data discretization and data resampling. The resampling algorithm is an iterative one which attempts to determine the best data to include/exclude in the training set and to determine the appropriate resampling rate. The appropriate resampling rate is a user-defined parameter which can be determined by computing the prediction accuracy for each resampling size. It is said appropriate if the increase in prediction accuracy started to stabilize at that point. Based on the analysed data, resampling rate generally depends on the gap of the majority and minority class. The smaller the gap between the majority and minority class means small increment size for resampling rate can be used. However, if the disparity is high, higher size increment is suggested.

Four datasets have been used, which are the crop yield dataset, credit risk dataset, ecoli dataset, and yeast dataset,

 ISBN: 978-1-61208-358-2

to test the performance of the framework. The prediction component of the framework attempts to investigate if the framework can be applied in any prediction/classification methodologies. Regression analysis, decision tree (DT), rule induction, and artificial neural networks (ANN) have been used as prediction/classification methodologies. The study reveals that the framework can be applied to any of the methodologies mentioned, though it works well in rule induction technique because it provides the highest overall and class accuracy.

The framework is also compared to the existing approaches in resolving class imbalance. The analysis reveals that the proposed framework is comparable to the hybrid technique, but the main difference is that the framework minimizes the oversampling and undersampling rate, but still gives good prediction accuracy.

For the future, there are different ways in which this study could be expanded. First, the procedure in determining the appropriate resampling size can be established. A mathematical model that can give the optimum resampling size can be done. Second, other performance measure, aside from prediction accuracy, can be investigated. Finally, prediction framework based algorithmic methodologies (internal approach) can also be studied.

VI. ACKNOWLEDGMENT

This research was funded by the Engineering Research and Development for Technology (ERDT) grant, under the Department of Science and Technology.

VII. REFERENCES

[1] Ling, C., and Li, C."Data Mining for Direct Marketing Problems and Solutions" The Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) New York, NY. AAAI Press, 1998

[2] Gorni, A. A., "The Application of Neural Networks in the Modeling of Plate Rolling Processes" 2008.

[3] Brennan, P. "A Comprehensive Survey of Methods for Overcoming the Class Imbalance Problem in Fraud Detection" 2012.

[4] Kubat, M., and Matwin, S. "Addressing the Curse of Imbalanced Training Sets: One Sided Selection" In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186 Nashville, Tennesse. Morgan Kaufmann, 1997.

[5] Hu, X. "Using Rough Sets Theory and Database Operations to Construct a Good Ensemble of Classifiers for Data Mining Applications," in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 233–240.

[6] Estabrooks, A., Jo, T., and Japkowicz, N., "A Multiple Resampling Method for Learning from Imbalanced Data Sets," Comput. Intell., vol. 20, no. 1, 2004, pp. 18–36.

[7] Patterson, L., "The Nine Most Common Data Mining Techniques Used in Predictive Analytics. 2010

[8] Japkowicz, N., and Stephen, S. "The Class Imbalance Problem: A Systematic Study." Intelligent Data Analysis 6:429–450, 2002.

[9] Chawla N. V., Bowyer K. W., Hall, L. O., and Kegelmeyer, W. P. "SMOTE: Synthetic Minority Over-sampling Technique" Journal of Artificial Intelligence Research 16:321– 357, 2002.

[10] Kaspar, T.C., et al., "Relationship Between Six Years of Corn Yields and Terrain Attributes," *Precision Agriculture*, 4(1), 2003, 87-101.

[11] Kumar, Ch. N., et.al., "An Updated Literature Review on the Problem of Class Imbalanced Learning in Clustering," in IJETR, Volume 2, Issue 2, February 2014.

[12] Barandela, R., Sánchez, J.S., García, V., and Rangel, E. "Strategies for Learning in Class Imbalance Problems" Pattern Recognition 36:849–851, 2003.

[13] Fernandez, A., Garcıa, S., del Jesus, M. J., and Herrera, F., "A Study of the Behaviour of Linguistic Fuzzy-Rule-Based Classification Systems in the Framework of Imbalanced Data-sets," Fuzzy Sets Syst., vol. 159, no. 18, 2008, pp. 2378–2398.

[14] KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3, 2011, 255-287.

[15] Accessed August 15, 2013. http://weka.sourceforge.net/packageMetaData/

A Study of VEPSO Approaches for Multiobjective Real World Applications

Omar Andres Carmona Cortes*, Andrew Rau-Chaplin†, Duane Wilson† and Jürgen Gaiser-Porter‡
*Informatics Academic Department
Instituto Federal de Educação, Ciência e Tecnologia do Maranhão
São Luis, MA, Brazil
omar@ifma.edu.br
†Risk Analytics Lab
Dalhousie University
Halifax, NS, Canada
arc@cs.dal.ca, dwilson@gmail.com
‡Global Analytics
Willis Group
London, UK
gaiserporterj@willis.com

***Abstract*—The purpose of this paper is to evaluate the performance of two approaches based on Vector Evaluated Particle Swarm Optimization (VEPSO) algorithm in two real world applications, which are the environmental economic dispatch problem and the optimization of a reinsurance contract portfolio. The two tested algorithms are the canonical VEPSO and a new version called VEPSO-N, where in the last one the global updating on each swarm is done based on the archive. The performance is evaluated using the following metrics: hypervolume, number of solutions and coverage, showing that both approaches can present good outcomes.**

Keywords–Real World Applications; Reinsurance Contract Optimization; Environmental Economic Dispatch; Vector Evaluated PSO; Multiobjective.

I. INTRODUCTION

Real world applications involve solving problems whose objectives are normally in conflict. For example, in an oil-based power plant, the lower the cost of generating energy, the bigger the emission of pollutants. Thus, companies have to figure out the best trade-off due to mainly environmental regulations. In a financial investment, the bigger the risk, the bigger the return. Therefore, investors are interested in smaller risks and to obtain bigger returns.

Those kind of problems are called Multiobjective Optimization Problems (MOPs) and their solution lay in the concept of Pareto Optimality, where solutions are characterized as a set of trade-off points. Gradient-based optimization techniques [1] [2] can be used to detect Pareto optimal solutions [3]; however, to do so the objectives have to be aggregated in a single objective function, and only one solution can be found per run [4], adding heavy computational cost to the whole process. In the same sense, traditional mono-objective evolutionary algorithms present the same kind of problem, *i.e.*, only one solution can be computed at the time.

Taking the computational cost into account, swarm/evolutionary multiobjective algorithms (MOEA) represent a viable alternative to solve MOPs. Indeed, MOEAs have been recognized to be well-suited for this kind of application because of their abilities to explore multiple solutions in parallel and to find a widespread set of non-dominated solutions in a single run [5]. Among the available MOEAs based on PSO is the VEPSO [6], which was based on Vector Evaluated Genetic Algorithm (VEGA) [7]. The idea behind VEPSO is to evolve two separated swarms, one per evaluation function, where considering two swarms, the direction of a swarm is guided by the best solution (g_{best}) found in the other one.

VEPSO has been successfully used in application ranging from the optimization of radiometry array antenna [8], passing through the optimization of a steady-state performance of a power system [9], to the optimization of the energy communications for heterogeneous network through cognitive sensing [10]. Its popularity can be mainly attributed to two reasons: it is easy to implement and the parallelization can be done in a straight-full way [4]. However, VEPSO might stagnate in bad approximation of the Pareto front as we can see in Matthysen et al. [11] and Lim et al. [12], where the authors show an analysis of VEPSO and VEGA, and explain why the canonical VEPSO tends to get trapped in sub-optimal Pareto frontiers.

In this context, a modification has been proposed by Lim et. al [13] in order to overcome problems with stagnation, where the best solution of a particular swarm is updated using a point stored in the archive, which contains only non-dominated solutions, rather than the g_{best} from the second swarm as in the canonical VEPSO algorithm. The main advantage of this approach is to keep the same number of callings to the evaluation function, therefore the comparison can be done in terms of iterations.

Two applications are considered in this paper. The first one is the Environmental Economic Dispatch problem (EED), where we want to determine the lower cost of generate energy using six generators and producing a smaller amount

 ISBN: 978-1-61208-358-2

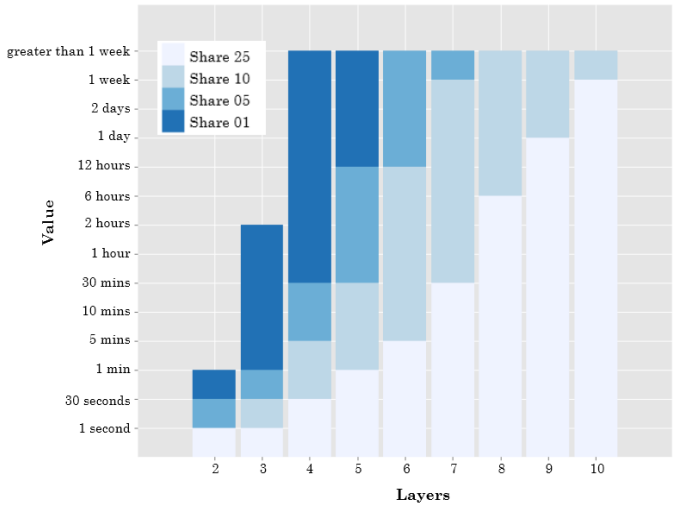


Figure 1. Estimated time for solving the RCO problem in R with different levels of discretization

of pollutants at the same time. The second application is the optimization of a Reinsurance Contract Portfolio (RCP) where, from the insurance company perspective, we want to hedge more risk and receive back more money in case of massive claims.

In terms of EED application, Qu's work [14] proposes a multiobjective fast-evolutionary programming for dealing with the problem; however, the Pareto frontiers are compared only visually, *i.e.*, he did not use proper metrics for comparison. Abedinia's work [15] deals with six and fourteen generators but the evaluation is done based only on the best results also without a proper metric. Farhat [16] did the same thing as previous works considering 3 different emission functions.

Regarding RCP, the first two works to address this kind of problem were [17] and [18]. Even though those two papers rely on mono-objective functions, they proved that it is worth to use evolutionary computation in this kind of application because it saves a considerable amount of time if compared with the enumeration method. Figure 1 shows the time required to solve the problem using the enumeration method, for instance, considering a 5% of discretization the enumeration method demands much more than a week to execute.

The remainder of this paper is organized as follows: Section II outlines the fundamentals of multiobjective problems; Section III presents the PSO-based algorithms, including VEPSO and the improved version called VEPSO-N; Section IV-A shows how the performance evaluation is done, the real world applications considered in this work and the simulation results; finally, Section V presents the conclusions and future works.

II. Multiobjective Fundamentals

A multiobjective optimization problem has to deal with two or more conflicting objective function [5] at the same time. These functions must be in conflict in order to build a Pareto frontier, where there are no solutions better than others; otherwise, the answer to the problem would be only one point in the search space.

Thus, assuming that a solution to a MOP is a vector in a search space X with m elements. A function $f : X \rightarrow Y$ evaluates the quality of a solutions mapping it into an objective space. Therefore, a multi-objective problem is defined as presented in (1), where f is a vector of objective functions, m is the dimension of the problem and n the number of objective functions.

$$Max\ y = f(x) = (f_1(x_1, ..., x_m), ..., f_n(x_1, ..., x_m)) \quad (1)$$

In order to determine whether a solution belongs to the Pareto frontier or not, we need the concept of optimality, which state that given two vectors $x, x* \in \Re$ and $x \neq x*$, x dominates $x*$ (denoted by $x \succeq x*$) if $f_i(x)$ is not worse than $f_i(x*), \forall\ i$ and $\exists$ at least one i where $f_i(x) > f_i(x*)$ in maximization cases and $f_i(x) < f_i(x*)$ otherwise. Hence, a solution x is said Pareto optimal if there is no solution that dominates x, in such case, x is called non-dominated solution. Mathematically, assuming a set of non-dominated solutions $\wp$, a Pareto frontier(pf) is represented as $pf = \{f_i(x) \in \Re | x \in \wp\}$

III. PSO-Based Algorithms

The particle swarm optimization was proposed by Kennedy and Eberhart [19] in 1995. The algorithm consists of particles which are placed into a search space, and move itself combining its own history position and the global optimal solution found so far. A particle position is represented in the search space as $X_i^D = (x_i^1, x_i^2, ..., x_i^D)$ and it is updated based on its velocity $V_i^D = (v_i^1, v_i^2, ..., v_i^D)$, where D represents the problem dimension. The new position is determined by means of (2) and (3), where w represents the inertia weight, c_l and c_2 are acceleration constants, r_l and r_2 are random number in the range $[0, 1]$, p_i^d is the best position reached by the particle P, and g^d is a vector which stores the global optima of the swarm.

$$v_i^d = w \times v_i^d + c_1 r_1 \times (p_i^d - x_i^d) + c_2 r_2 \times (g^d - x_i^d) \quad (2)$$

$$x_i^d = x_i^d + v_i^d \quad (3)$$

The Algorithm 1 outlines how PSO works. Initially, the swarm is created at random, where each particle has to be within the domain $[a_i^d, b_i^d]$. Then particles are evaluated in order to initialize the P matrix and the g^d vector, which are the best experience of each particle and the best solution that has been found so far, respectively. Thereafter, the velocity and the position of a particle are updated within a loop that obeys some stop criteria. In the pseudo code presented in the Algorithm 1, the stop criteria is a certain number of iterations.

A. VEPSO

The VEPSO algorithm is a multiobjective heuristic based on Vector Evaluated Genetic Algorithms (VEGA) [7]. The main idea behind this algorithm is to "evolve" two independent swarms and exchange information between them, *i.e.*, assuming two swarms S_1 and S_2, and two functions to be

 ISBN: 978-1-61208-358-2

Generate a swarm of particles **X** of size *s* from $[a_i^d, b_i^d]$;
for *i = 1* to *swarm_size* **do**
  Evaluate swarm
  Update the best position g
  Update p of the particles
  for *j = 1 to D* **do**
    Update velocity V using (2)
    Update position X using (3)
  end
end
Verify if the current g is better than the best of the current swarm

Algorithm 1: Particle Swarm Optimization (PSO)

optimized f_1 and f_2, being solved by S_1 and S_2, respectively. The swarm S_1 updates its velocity using the best particle of S_2 (g_2). On the other hand, S_2 updates its velocity based on the best particle of S_1 (g_1). Then, an archive is held after each iteration joining both swarms in order to obtain only the non-dominated solutions.

B. VEPSO-N

The VEPSO-N uses the idea behind Lim's work [13] where an archive with non-dominated solutions is maintained, then each swarm updates its own global best (g_{best}) using the best result in its respective function. Mathematically, considering two swarms S_1 and S_2, and $A = \{a_1, a_2, ..., a_n\}$ as being the set of non-dominated solutions, if $f(a_i)$ presents the best solutions regarded to S_1 and $f(a_j)$ shows the best solutions regarded to S_2, then a_i and a_j replace $g_{best}^{S_1}$ and $g_{best}^{S_2}$, respectively. In other words, we use the best solutions in the archive for updating the global best on each swarm.

IV. Experiments

A. Performance Evaluation and Parameters

In this section, we discuss the experimental evaluation of EED and RCP optimization problems as follows. Firstly, the number of non-dominated points found in the Pareto frontier after 31 trials were determined. Secondly, the hypervolume, which is the volume of the dominated portion of the objective space, as presented in (4), was measured, where for each solution $i \in Q$ a hypercube v_i is constructed. Having computed each v_i, we can calculate the final hypervolume by the union of all of them.

$$hv = volume(\bigcup_{i=1}^{|Q|} v_i) \tag{4}$$

Thirdly, the dominance relationship between Pareto frontiers (coverage) obtained with different algorithms was calculated as depicted in (5). Roughly speaking, the coverage is the ratio between the number of solutions dominated by A divided by the number of elements from set B [27]. If $C(A, B) = 1$ then all solutions in A dominate B. Therefore, $C(A, B) = 0$ means the opposite.

$$C(A, B) = \frac{|\{b \in B | \exists a \in A : a \preceq b\}|}{|B|} \tag{5}$$

The parameters used in all experiments for PSO algorithms were: $c_1 = c_2 = 0.5 + log(2)$; numbers of particles = 50; w_0 = 0.9; w_f = 0.1, where w has a linear updating based on (6) as proposed by Nikabadi and Ebadzadeh [22], where w_0 is the initial weight, w_f is the final one, N_G is the number of iterations and i depicts the current generations. Moreover, all study cases are executed using 500, 1000 and 2000 iterations; the initialization was done as recommended by Clerc [23]; and, we are not narrowing the archive size.

$$w = (w_0 - w_f) \times \frac{N_G}{i} \tag{6}$$

All the parameters were chosen empirically and all tests have been conducted using R version 2.15.0 [24] and RStudio [25] on a Windows 7 64-bit Operating System running on an Intel i7 3.4 Ghz processor, with 16 GB of RAM.

B. Real World Applications

1) Environmental Economic Dispatch: The environmental economic dispatch involves the optimization of both fuel cost and pollution emission simultaneously as presented in (7) and (8), where P_i is the power used on the i^{th} generator, a_i, b_i, c_i, α_i, β_i and γ_i are coefficients presented in Tables I and II.

$$min\ Fc = \sum_{i=1}^{n}(a_i P_i^2 + b_i P_i + c_i) \tag{7}$$

$$min\ E = \sum_{i=1}^{n}(\alpha_i P_i^2 + \beta_i P_i + \gamma_i) \tag{8}$$

subject to

$$\sum_{min}^{max} P_i \geq P_d \tag{9}$$

$$P_{min} \leq P_i \leq P_{max} \tag{10}$$

The constraint presented in (9) represents the required demand, *i.e.*, the sum of all powers has to be equal or greater than a specific demand, and the constraint shown in (10) depicts the operation boundaries of each generator which are also presented in Tables I and II. The coefficients and boundaries where obtained from Singh and Kumar [26], and, at this stage, we are not considering the power loss. Moreover, we are taking into account demands equals to 500 MW and 700 MW.

TABLE I. Generators and cost coefficients

P_{min}	P_{max}	a	b	c
5	50	0.01	2	10
5	60	0.012	1.5	10
5	100	0.004	1.8	20
5	120	0.006	1	10
5	100	0.004	1.8	20
5	60	0.01	1.5	10

Table III shows the average results in terms of number of solutions and hypervolume for a EED problem comprises of 6

TABLE II. GENERATORS AND EMISSION COEFFICIENTS

P_{min}	P_{max}	α	β	γ
5	50	0.00419	0.32767	13.85932
5	60	0.00419	0.32767	13.85932
5	100	0.00683	-0.54551 4	0.26690
5	120	0.00683	-0.54551	40.26690
5	100	0.00461	-0.51116	42.89553
5	60	0.00461	-0.51116	42.89553

generators and a demand of 500 MW, following the legend I for VEPSO and II for VEPSO-N. As expected, increasing the number of iterations the number of non-dominated solutions also increases in both algorithms. We can also observe the number of solutions in VEPSO-N is smaller than in VEPSO. On the other hand, the average hypervolume is better in VEPSO-N, showing that the Pareto frontier might be better on the respective algorithm.

TABLE III. NUMBER OF SOLUTIONS AND HYPERVOLUME FOR EED PROBLEM CONSIDERING A DEMAND OF 500 MW

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
#NS	67.64	57.61	177.13	62.93	399.42	74.71
HV	1.8e6	1.96e6	1.91e6	2.46e6	1.74e6	2.09e6

Table IV shows the coverage metrics in the final Pareto frontier, which indicate that VEPSO presents better solutions, particularly with 2000 iterations, where VEPSO dominates 74% of the solutions. Figure 2 shows the final Pareto frontier after 31 trials, where we can see that VEPSO tends to find out a better Pareto frontier. When 2000 iterations are used, VEPSO-N presents an extension of the Pareto frontier; even though, those points are not interesting because they are dominated points. Also, we have to point out that VEPSO-N tends to concentrate its solutions in the beginning of the Pareto frontier when 500 iterations are used, which is not good in terms of diversity.

TABLE IV. COVERAGE FOR EED PROBLEM WITH DEMAND OF 500 MW

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
I	-	0.25	-	0.28	-	0.74
II	0.07	-	0.23	-	0.25	-

Table V presents the average results in terms of number of solutions and hypervolume for a EED problem comprises of 6 generators and a demand of 700 MW, following the legend I for VEPSO and II for VEPSO-N. The problem is harder to be solved because the demand constraint is stronger. Nonetheless, the number of solutions increase as we increase the number of iterations. However, this increment is clearly bigger in VEPSO than in VEPSO-N. On the other hand, the hypervolume in VEPSO-N is better indicating that its solutions might be slightly better. Figure 3 shows the final Pareto frontier after 31 trials. Visually, it seems that the final frontier is better for VEPSO which is confirmed in Table VI, excepting for 500 iterations where VEPSO-N dominates 36% of the solutions, whereas VEPSO dominates 29%.

Summarizing, VEPSO seems to cover better the final Pareto frontier than VEPSO-N for demands of 500 and 700 MW. On the other hand, when the demand constraint is harder (700 MW), VEPSO-N presented best results with lower number of iterations.

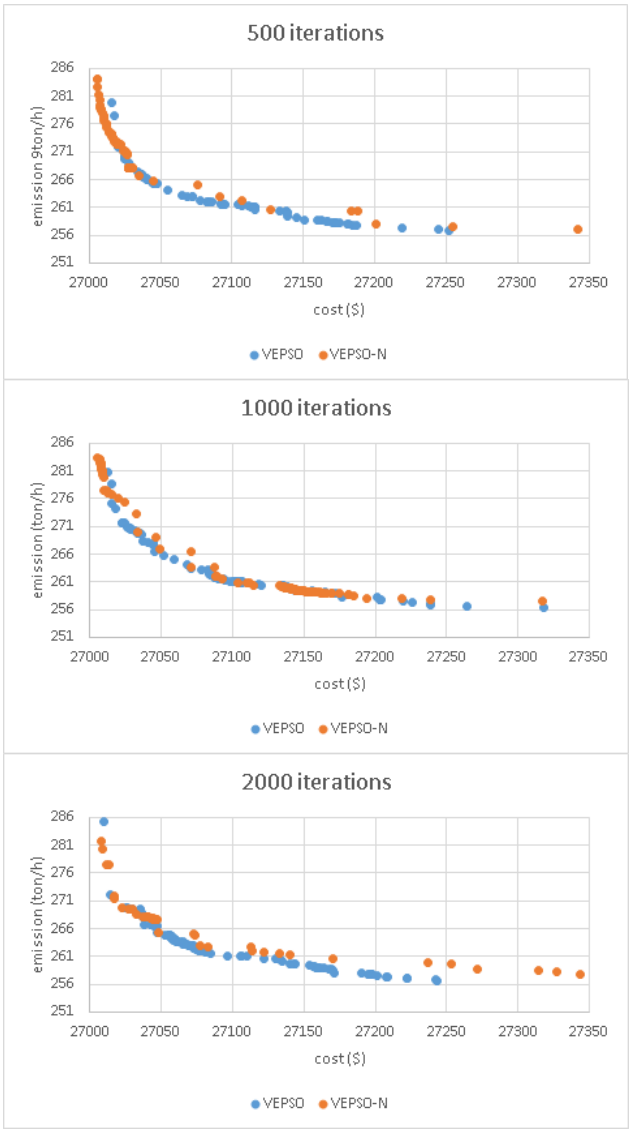


Figure 2. Pareto frontier after 31 executions for EED and demand of 500

TABLE V. NUMBER OF SOLUTIONS AND HYPERVOLUME FOR EED PROBLEM CONSIDERING A DEMAND OF 700 MW

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
#NS	141.9	93.6	323.81	103.94	546.58	137.42
HV	4.8e6	4.9e6	4.8e6	5.3e6	5.3e6	5.4e6

TABLE VI. COVERAGE FOR EED PROBLEM WITH DEMAND OF 700 MW

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
I	-	0.29	-	0.35	-	0.37
II	0.36	-	0.30	-	0.13	-

2) Reinsurance Contract Optimization: The reinsurance process consists of hedging risk from the insurance company to a bigger one, called reinsurance company. The main purpose of doing so is to survive in case of massive claims mainly caused by natural catastrophes. The reinsurance contract optimization problem consists of given a treaty structure to figure out the best combination of placements or shares in order to transfer

 ISBN: 978-1-61208-358-2

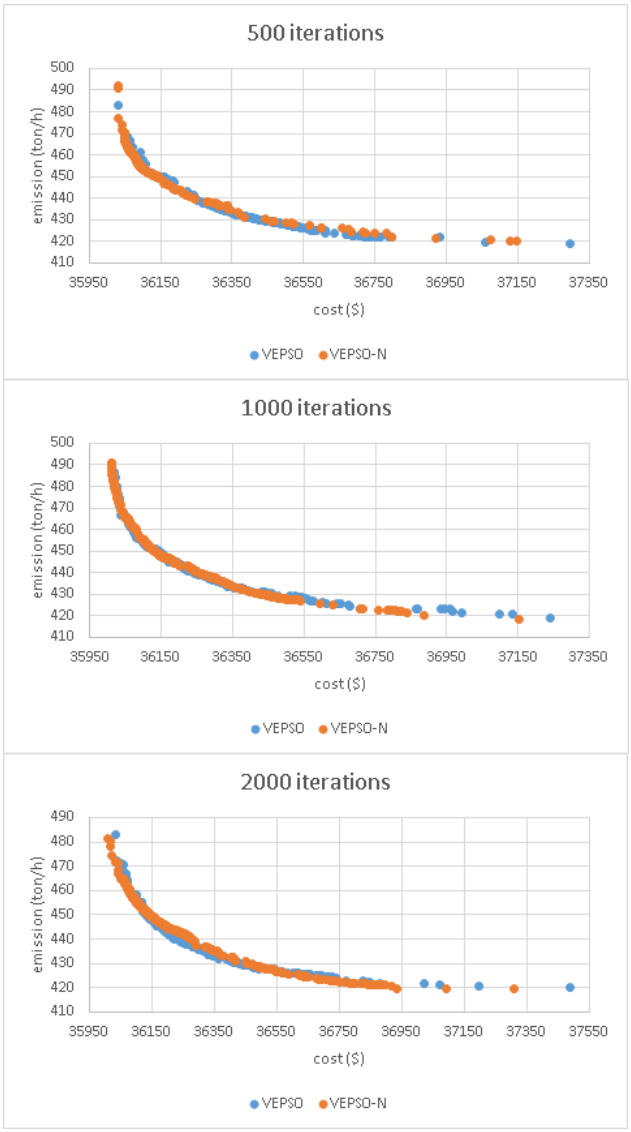


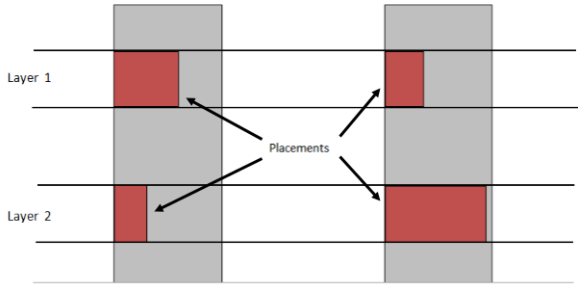
Figure 3. Pareto frontier after 31 executions for EED and demand of 700

the maximum of risk, and at the same time, to receive the maximum return when facing massive claims. Therefore, the main purpose of a RCO problem is to find out the best combination of shares or placements which maximize both the transfered risk and the expected return. Figure 4 is an example of a structure with two different solutions in terms of shares.

The Equation 11 represents the RCO in terms of an optimization problem, where VaR is a risk metric, $\mathbf{R}$ is a function based on a combination of shares pi, and E is the expected value. For further details about the problem refer to [17] and [20].

$$\begin{aligned} &maximize && f_1(x) = VaR_\alpha(\mathbf{R}(\pi)) \\ &maximize && f_2(x) = E[\mathbf{R}(\pi)] \end{aligned} \tag{11}$$

Table VII shows the average number of solutions and the average hypervolume (all hypervolume values are multiplied by 1×10^{15}) for the RCO problem consisting of 7 layers of real anonymized data and a discretization of 5% obtained by



Solution	Layer	Limit	Deductible	Placement	Max Recovery	Premium
1	Layer 1	200	300	60%	120	$40
	Layer 2	200	100	20%	40	$40
2	Layer 1	200	300	30%	60	$25
	Layer 2	200	100	95%	190	$100

Figure 4. Structure and two solutions with different placements

rounding solutions. The following legends are I for VEPSO and II for VEPSO-N. Again, the number of solutions increase as we increase the number of iterations for both approaches; however, in this particular experiment, VEPSO-N presented a greater number of solutions for 500 and 2000 iterations, nonetheless the average hypervolumes were smaller, indicating that VEPSO could present better solutions.

TABLE VII. NUMBER OF SOLUTIONS AND HYPERVOLUME FOR RCO PROBLEM WITH 7 LAYERS

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
#NS	1394	2152.52	4677.13	4360.06	11650	16360.65
HV	2.26	2.21	2.26	2.2	2.26	2.26

Figure 5 depicts the final Pareto frontier for RCO problem using 7 layers for 500, 1000 and 2000 iterations, respectively. Visually, solutions seem to be very similar, perhaps with a little advantage to VEPSO, specially when 2000 iterations are taking into account. Thus, Table VIII presents the coverage metrics on the final Pareto frontier, where VEPSO dominates 4% more solutions with 500 iterations, and gets worse with 1000 iterations dominating 6% less solutions. This scenario is confirmed with 2000 iterations where VEPSO-N could dominate even more points, *i.e.*, 30%.

TABLE VIII. COVERAGE FOR RCO PROBLEM WITH 7 LAYERS

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
I	-	0.31	-	0.21	-	0.17
II	0.27	-	0.27	-	0.30	-

In order to increase the difficulty of solving this optimization problem, 8 layers were synthetically added to the previous 7 layers structure. Table IX shows the results in terms of the average number of solutions and average hypervolume (all hypervolume values are multiplied by 1×10^{15}) when 15 layers are considered. Again, the number of solutions increase as the number of iterations; however, in this particular case, VEPSO and VEPSO-N presented similar results, excepting for 2000 iterations where VEPSO reached more non-dominated points. On the other hand, VEPSO presented better average hypervolume, indicating better solutions.

Figure 6 represents the final Pareto frontier for RCO problem using 15 layers for 500, 1000 and 2000 iterations,

 ISBN: 978-1-61208-358-2

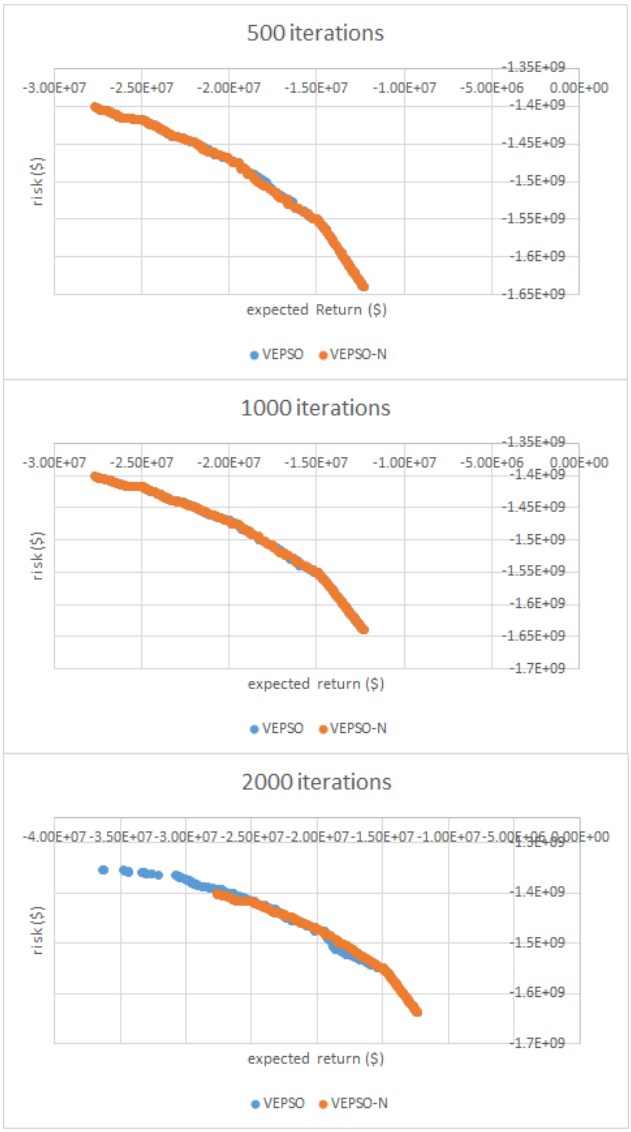


Figure 5. Reinsurance contract optimization Pareto frontier after 31 executions using 7 layers

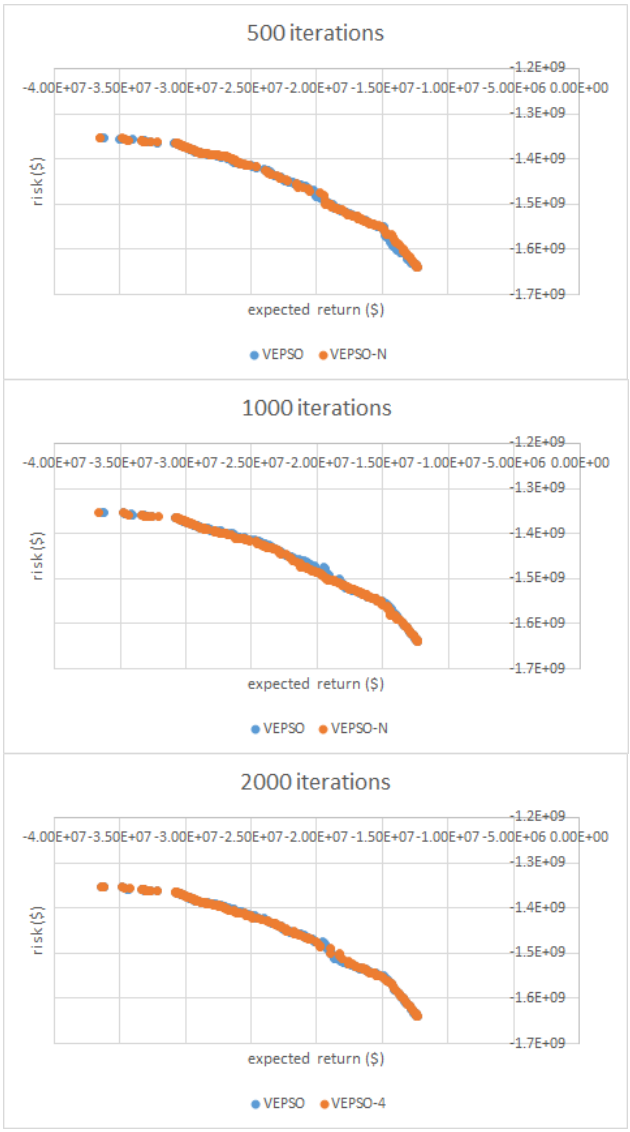


Figure 6. Reinsurance contract optimization Pareto frontier after 31 executions using 15 layers

TABLE IX. NUMBER OF SOLUTIONS AND HYPERVOLUME FOR RCO PROBLEM WITH 15 LAYERS

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
#NS	186.68	453.9	1001.97	2214.1	3916.8	3299.7
HV	4.57	4.1	4.6	4.1	4.55	4.2

TABLE X. COVERAGE FOR RCO PROBLEM WITH 15 LAYERS

	500 it		1000 it		2000 it	
	I	II	I	II	I	II
I	-	0.23	-	0.56	-	0.44
II	0.55	-	0.22	-	0.33	-

respectively. Solving the problem with more layers is clearly more difficult. Nonetheless, both algorithms presents visually similar results. Thus, Table X shows the coverage for the algorithms where we can see that differences in this particular case are evident in the following cases: (i) VEPSO-N with 500 iterations dominating 55% of VEPSO solutions; (ii) VEPSO with 1000 iterations dominating 56% of VEPSO-N points; and, VEPSO with 2000 iterations dominating 44% of VEPSO against 33% in VEPSO-N.

V. CONCLUSION

This paper presented a study about the performance of the canonical VEPSO and the algorithm called VEPSO-N in solving multiobjective world real problems. The comparison shows that both approaches are suitable for finding out non-dominated points; however, the traditional VEPSO has an advantage in the EED problems, since present more solutions in the search space and tends to dominate more points as well, for example, using 2000 iterations VEPSO domains 70% of the solutions from VEPSON-N with a demand of 500 MW against 25% in the way around. Further, when a demand of 700 MW is considered, VEPSO domains 37% of the solutions from VEPSO-N using 2000 iterations.

 ISBN: 978-1-61208-358-2

In RCO, VEPSO-N tends to present more solutions than VEPSO with less layers, and dominates up to 30% of the solutions from VEPSO. This behavior is similar when solving the problem with 15 layers using 500 and 1000 iterations; nevertheless, VEPSO-N is overcame when 2000 iterations are used, since VEPSO starts to domain 44% of points against %33 of VPSON-N.

Future work includes a comparison against other modern approaches, such as Vector Evaluated Differential Evolution (VEDE) [28], Strength Pareto Evolutionary Algorithm (SPEA2) [29] and Multiobjective Evolutionary Algorithm/Distributed (MOEA/D) [27], hybridization of VEPSO with other metaheuristics and a parallel version.

ACKNOWLEDGMENT

This research was financially supported by Flagstone Re, Halifax, Canada and by the Science without Border program of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil), and Instituto Federal de Educação, Ciência e Tecnologia do Maranhão. We also would like to thank the Willis group for providing the anonymized real world data for the RCO problem.

REFERENCES

[1] Y. Bengio, "Gradient-Based Optimization of Hyperparameters", Neural Computation , vol. 12, no. 8, 2000, pp. 1889–1900.

[2] R. Haupt, "Comparison Between Genetic and Gradient-Based Optimization Algorithms for Solving Electromagnetics Problems", IEEE Transaction on Magnetics, vol. 31, no. 03, 1995, pp. 1932–1935.

[3] D. A. Iancu and N. Trichakis, "Pareto Efficiency in Robust Optimization", Articles in Advance Management Science, 2013, pp. 1–18.

[4] K. E. Parsopoulos, D. K. Tasoulis, M. N. Vrahatis, and K. Words, "Multiobjective Optimization Using Parallel Vector Evaluated Particle Swarm Optimization", In Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, pp. 823–828, ACTA Press, 2004.

[5] K. Deb, "Multi-objective Optimization using Evolutionary Algorithms", John Wiley and Sons LTDA, 2001.

[6] K. E. Parsopoulos and M. N. Vrahatis, "Particle Swarm Optimization Method in Multiobjective Problems", In Proceedings of the 2002 ACM Symposium on Applied Computing, pp. 603–607, ACM Press, 2002.

[7] J. Schaffer, "Multiple objective optimization with vector evaluated genetic algorithms", in Proceedings of the 1st International Conference on Genetic Algorithms, pp. 93–100, L. Erlbaum Associates Inc. Hillsdale, NJ, USA, 1985.

[8] D. Glies and Y. Rahmat-Samii, "Vector evaluated particle swarm optimization (VEPSO): Optimization of a radiometer array antenna", in Proc. of the IEEE International Symposium on Antennas and Propagation,vol. 3, 2004, pp. 2297-2300.

[9] J. G. Vlachogiannis and K. Y. Lee, "Multi-objective based on parallel vector evaluated particle swarm optimization for optimal steady-state performance of power systems", Expert Systems with Applications, vol. 36, no. 8, 2009, pp. 802-808.

[10] S. Hou, X. Zhang, H. Zheng, L. Zhao, and W. Fang, "An effective interference management framework to achieve energy-efficient communications for heterogeneous network through cognitive sensing", International ICST Conference on Communications and Networking in China (CHINACOM), pp. 536,541, 2012.

[11] W. Matthysen, A. P. Engelbrecht, and K. M. Malan, "Analysis of stagnation behavior of vector evaluated particle swarm optimization", IEEE Symposium on Swarm Intelligence (SIS), pp. 155,163, 2013

[12] K. S. Lim et al., "Convergence and diversity evaluation for vector evaluated Particle Swarm Optimization", Proceedings of International Conference on Modelling, Identification & Control (ICMIC), pp. 280-285, 2013.

[13] K. S. Lim et al., "Improving Vector Evaluated Particle Swarm Optimisation by Incorporating Nondominated Solutions", The Scientific World Journal, vol. 2013, Article ID 510763, 2013.

[14] B.-Y. Qu, P. N. Suganthan, V. R. Pandi, and K. B. Panigrahi, "Multi objective evolutionary programming to solve environmental economic dispatch problem", International Conference on Control Automation Robotics & Vision (ICARCV), pp. 1673-1679, 2010.

[15] O. Abedinia, N. Amjady, and M. S. Naderi, "Multi-objective Environmental/Economic Dispatch using firefly technique", International Conference on Environment and Electrical Engineering (EEEIC), pp. 461-466, 2012.

[16] I. A. Farhat and M. E. El-Hawary, "Multi-objective economic-emission optimal load dispatch using bacterial foraging algorithm", IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), pp. 1-5, 2012.

[17] O. A. C. Cortes, A. Rau-Chaplin, D. Wilson, and J. Gaiser-Porterz, "Efficient Optimization of Reinsurance Contracts using Discretized PBIL", In Proceedings of Data Analytics, pp. 18-24, Porto, 2013.

[18] O. A. C. Cortes, A. Rau-Chaplin, D. Wilson, and J. Gaiser-Porter, "On PBIL, DE and PSO for Optimization ofReinsurance Contracts", EvoStar, EvoFin, Barcelona, 2014.

[19] J. Kennedy and R. Eberhart, "Particle swarm optimization", Proceedings of IEEE International Conference on Neural Networks, vol. 4, pp. 1942-1948, 1995.

[20] J. Cai, K. S. Tan, C. Weng, and Y. Zhang, "Optimal reinsurance under VaR and CTE risk measures", Insurance: Mathematics and Economics, no. 43, 2008, pp. 185-196.

[21] R. Giusti and G. E. A. P. A. Batista, "Discovering Knowledge Rules with Multi-Objective Evolutionary Computing", 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 119–124, 2010.

[22] A. Nikabadi and M. Ebadzadeh, "Particle swarm optimization algorithms with adaptive Inertia Weight : A survey of the state of the art and a Novel method", IEEE journal of evolutionary computation, 2008.

[23] M. Clerc, "Standard Particle Swarm Optimisation", hal-00764996, version 1, Retrieved [June,2013], available in: http://hal.archives-ouvertes.fr/hal-00764996, 2012.

[24] _, "The Comprehensive R Archive Network", Retrieved [August, 2014], available in: http://cran.r-project.org/, 2014.

[25] _, "RStudio", Retrieved [August, 2014], available in: http://www.rstudio.com/, 2014.

[26] N. Singh and Y. Kumar, "Economic load dispatch with environmental emission using MRPSO", IEEE 3rd International Advance Computing Conference (IACC), pp. 995-999, 2013.

[27] Q. Zhang and H.i Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition", IEEE Transactions on Evolutionary Computation, vol.11, no.6, 2007, pp. 712–731.

[28] K. E. Parsopoulos, D. K. Tasoulis, N. G. Pavlidis, V. P. Plagianakos, and M. N. Vrahatis, "Vector Evaluated Differential Evolution for Multiobjective Optimization", Congress on Evolutionary Computation, vol. 1, pp. 204–211, 2004.

[29] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the Strength Pareto Evolutionary Algorithm", Technical Report, Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich, 2001.

 ISBN: 978-1-61208-358-2

Application of Change-Point Detection in Image Retrieval

Dongwei Wei
and Yuehua Wu

Department of Mathematics and Statistics
York University,
Toronto, Ontario, Canada
Email: weidw@mathstat.yorku.ca, wuyh@mathstat.yorku.ca

Xiaoping Shi

Department of Mathematics,
Statistics and Computer Science
St. Francis University,
Antigonish, Nova Scotia, Canada
Email: shermanshixp@gmail.com

***Abstract*—In this paper, we approach the image retrieval problem from a different angle by converting the problem into a change-point detection problem. An algorithm is introduced for detecting multiple change-points in distributions of a sequence of independently distributed random variables. By using this algorithm, a procedure is given for image retrieval. The proposed method is evaluated via two examples, which show that the proposed method for the image retrieval has satisfactory performance in terms of the quality of the retrieved images.**

Keywords–Change-point detection; Image retrieval; Contamination; Outliers

I. INTRODUCTION

The term "change-point" refers to a time moment or a spatial location at which the data generation process undergoes an abrupt change so that a different model needs to be used to characterize the generation mechanism after the change by [13]. Statistical studies of change-point problems started with [12] and have flourished especially since the 1980s (see the books [2] and [1] among others). Results of these studies have found applications in a wide range of areas such as quality control, finance, environmetrics, medicine, geographics, and engineering. Statistical models in change-point problems may vary in different application areas. The one used in this paper is the change-point in probability distribution, referring to a generic change of the distribution of observations before and after the change-point. Another popular one is the change-point in linear regression which includes the change-point in mean as its special case. As commented in [13], the essential difference between the model with change-points and the usual piecewise model is that the points of change in the latter are specified while in the former they are unknown and need to be estimated. In addition, for general change-point models, it is unknown whether change-points even exist, and when they exist, how many there are. This uncertainty adds to the difficulty and complexity in analyzing change-point models. Therefore detecting all change-points in a data series has become an important task in the analysis of change-points. It is well known that if there exist change-points, it is not appropriate to make a statistical analysis without considering their existence and the results derived from such an analysis may be misleading.

In digital image analysis, each grey-scale image consists of a number of pixels that are elements of a matrix (named as image matrix hereafter) (see [7] and [10] among others). Dimensions of the image matrix for a fine image are usually very large. If an image is corrupted, and hence the corresponding image matrix is no longer the one for the original image, the challenging problem is how to retrieve the original image or equivalently recover its image matrix. There is a rich literature on image retrieval which includes [3] [7] [8] [11] [14] among others. In this paper, we approach the image retrieval problem from a different angle. By considering each row and each column of the image matrix of this corrupted image as data sequences, we can convert the image retrieval problem into a multiple change-point detection problem. Thus the image retrieval problem may be solved by employing a multiple change-point detection method.

The paper is arranged as follows. In Section 2, we briefly introduce the problem of multiple change-points in distributions and then present the multiple change-point detection algorithm proposed in [16]. In Section 3, we give a procedure for image retrieval. In Section 4, we evaluate the performance of the proposed method via two examples. We complete this paper with some concluding remarks in Section 5.

II. MULTIPLE CHANGE-POINT DETECTION

Let $X_1, \cdots, X_n$ be independently distributed random variables with distributions $F_i, i = 1, 2, \cdots, n$, respectively. The problem of multiple change-points in distributions is that there exist $1 < k_1 < k_2 < \cdots < k_p < n$ such that $F_1 = \cdots = F_{k_1} \neq F_{k_1+1} = \cdots = F_{k_2} \neq F_{k_2+1} = \cdots = F_{k_3} \neq \cdots \neq F_{k_{p-1}+1} = \cdots = F_{k_p} \neq F_{k_p+1} = \cdots = F_n$. The task of multiple change-point detection is to find the number of change-points, i.e., to find p, and to determine the locations of these change-points, i.e., to estimate $k_1 < k_2 < \cdots < k_p$. Some work on detecting a change in distribution include [5], [6], [4], and [16] among others. The following material is mainly based on [16].

First, consider the single change-point detection problem. One needs to find out if there exists a change in distribution at k^* such that $F_1 = \cdots = F_{k^*} \neq F_{k^*+1} = \cdots = F_n$ with $1 < k^* < n$. It is noted that k^* is unknown. If $k^* = 1$ or n, we consider that there is no change in distribution.

Let

$$C_k(t) = \sqrt{\frac{k(n-k)}{n}} \left(\frac{1}{k} \sum_{i=1}^{k} \cos(tX_i) - \frac{1}{n-k} \sum_{i=k+1}^{n} \cos(tX_i) \right), \quad (1)$$

which is based on the real part of empirical characteristic function combining with the traditional cumulative sum method. If

 ISBN: 978-1-61208-358-2

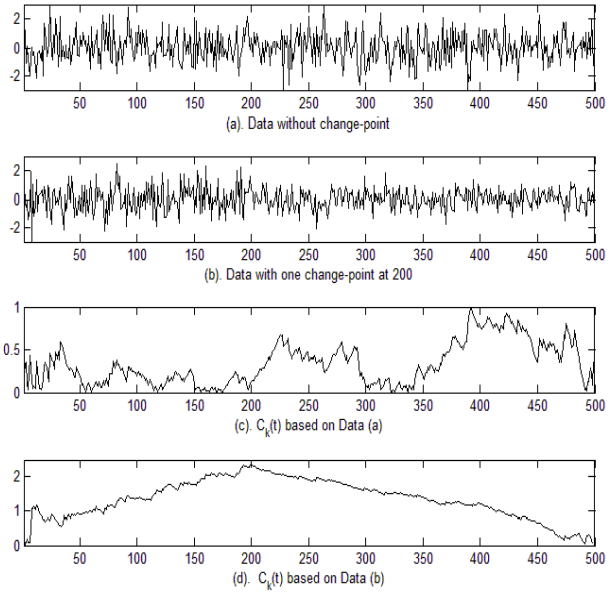


Figure 1. Top two panels: (a) data without any change in distribution; (b) data with one change in distribution at $k^* = 200$. Bottom two panels: plots of $C_k(t)$ with $t = 0.2$ based on the data displayed respectively in (a) and (b).

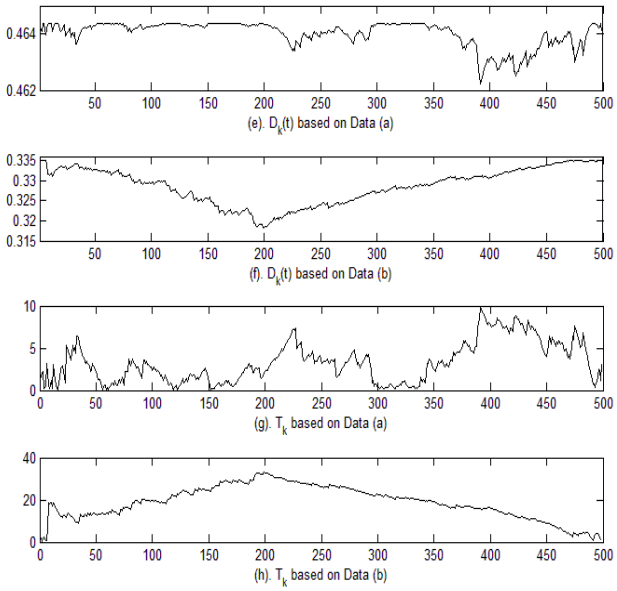


Figure 2. Top two panels: plots of $D_k(t)$ with $t = 0.2$ based on the data displayed respectively in (a) and (b) of Figure 1. Bottom two panels: plots of T_k based on the data displayed respectively in (a) and (b) of Figure 1.

k is the true change-point, i.e., $k = k^*$, the value of $|C_k(t)|$ is expected to be the largest.

Define

$$T_k = \int_t \omega(t) \frac{C_k(t)}{D_k(t)} dt, \tag{2}$$

where $\omega(t) = \frac{2}{\pi}\sqrt{(1-t^2)}$, $t \in [-1, 1]$,

$$D_k^2(t) = \frac{1}{n}\Big\{ \sum_{i=1}^{k} \Big(\cos(tX_i) - \frac{1}{k}\sum_{j=1}^{k} \cos(tX_j) \Big)^2 + \sum_{i=k+1}^{n} \Big(\cos(tX_i) - \frac{1}{n-k}\sum_{j=k+1}^{n} \cos(tX_j) \Big)^2 \Big\} \tag{3}$$

Illustrations of the statistics $C_k(t)$, $D_k(t)$ and T_k are plotted in Figures 1-2, where the true change-point is located at $k^* = 200$ and the sample size $n = 500$. For the case that there is no change-point, $F_1 = \cdots = F_n = N(0, 1)$, while for the case that there is a change-point, $F_1 = \cdots = F_k = N(0, 1)$ and $F_{k+1} = \cdots = F_n = N(0, 0.36)$.

It is shown in [16] that if $X_1, \cdots X_n$ are independent identically distributed random variables, under the assumption that there is no change in distribution, then

$$\lim_{n\to\infty} P\{A(\log n) \max_{1\le k\le n} T_k \le u + D(\log n)\} = \exp(-2e^{-u}) \tag{4}$$

where $A(x) = (2\log x)^{1/2}$ and $D(x) = 2\log x + 0.5\log\log x - 0.5\log\pi$.

To perform change-point detection, we first calculate the statistic $\max_k T_k$ and then compare it with the corresponding critical value which is computed by using (4). If $\max_k T_k$ is smaller than the critical value, no change-point is claimed, otherwise there exists a change-point which is estimated by $\hat{k} = \arg\max_{1<k<n} |T_k|$.

Now consider the multiple change-point detection problem introduced in the beginning of this section. In light of the iterated cumulative sums of squares algorithm in [9], [16] proposed an efficient and fast algorithm for detecting multiple change-points in distributions, which is given below. This algorithm will be used in our image retrieval procedure given in Section 4.

Let $X[l_1 : l_2]$ represent the segment $X_{l_1}, X_{l_1+1} \cdots, X_{l_2}$ with $l_1 < l_2$. Denote $T_k(t)$ that is computed in terms of $X[l_1 : l_2]$ by $T_k(X[l_1 : l_2])$. Define $k^*(X[l_1 : l_2])$ to be the point at which $M(X[l_1 : l_2]) \equiv \max_k T_k(X[l_1 : l_2])$ is attained. Let $CV(X[l_1 : l_2])$ be the critical value computed via (4). Denote the set of detected change-points by CP. The pseudo-code of the algorithm is as follows:

1) Set $l_1 = 1$ and $l_2 = n$. Calculate $M(X[l_1 : l_2])$ and $k^*(X[l_1 : l_2])$;
2) While $M(X[l_1 : l_2]) > CV(X[l_1 : l_2])$, repeat 3) − 12);
3) Set $k_{first} = k_{last} = k^*(X[l_1 : l_2])$;
4) Set $M_1 = M(X[l_1 : k_{first}])$ and $k_1 = k^*(X[l_1 : k_{first}])$;
5) While $M_1 > CV(X[l_1 : k_{first}])$, repeat 6);
6) $k_{first} = k_1; M_1 = M(X[l_1 : k_{first}])$;
7) Then set $M_2 = M(X[k_{last} : l_2])$ and $k_2 = k^*(X[k_{last} : l_2])$;
8) While$(M_2 > CV(X[k_{last} : l_2]))$, repeat 9);
9) $k_{last} = k_2; M_2 = M(X[k_{last} : l_2])$;
10) If $k_{first} == k_{last}$, there is only one change-point in $[l_1 : l_2]$, add it in CP and end the loop;
11) Else add the two candidate change-points in CP and then continue;

12) Reset $l_1 = k_{first}, l_2 = k_{last}$;

13) Sort change-point in CP. Denote the size of CP by N. Add 0 and n in CP. Then $CP = \{\ell_0 = 0, \ell_1, \ldots, \ell_N, \ell_{N+1} = n\}$ and $\ell_i < \ell_j$ if $i < j$.
14) For $j = 1, \ldots, N$, check whether a possible change-point exists between $[\ell_{j-1} + 1 : \ell_{j+1}]$, if yes, keep the jth change-point; if not, eliminate the jth change-point from CP.
15) Repeat $14)$ until CP does not change.

We name the above algorithm as *Algorithm WSW*. Here "WSW" is the acronym of the last names of the three authors of [16]. By applying this algorithm, one can not only estimate the number of multiple change-points (if the number is 0, there is no change in distribution) but also estimate multiple change-point locations.

III. IMAGE RETRIEVAL

First let us consider a noise contaminated black white image with scratches. If we treat each row (column) of the image matrix of this corrupted image as a data series, it is easy to see that each data series contains change-points that reflect the color changes. We remark that the locations of scratches do not correspond to locations of change-points, instead, they correspond to the locations of outliers. Thus a robust statistical change-point detection method may be applied to locate the true change-points, and hence the image matrix for the original image can be recovered. Since there are large number of rows even for a small image, it is important to employ a fast and efficient change-point detection method to tackle such a problem. *Algorithm WSW* (see the previous section) is suitable for this task.

For each row (column) of the image matrix, we employ *Algorithm WSW* to find out whether or not there exist multiple change-points and then estimate their locations if multiple change-points do exist. These multiple change-points are used to segment the data sequence in each row (column), which is the key in our image retrieval method. We are now ready to propose a procedure for image retrieval, which consists of the following steps:

1) Convert the noise contaminated black white image into the image matrix, denoted by A.
2) Use *Algorithm WSW* to detect multiple change-points for each row of A that divide the data sequence in each row into segments.
3) For each row of A, replace all the numbers within each segment by the segment median. Denote the resulting matrix by $A_r = (a_{ij}^{(r)})$.
4) Repeat steps 2) and 3) for each column of the matrix A. Denote the resulting matrix by $A_c = (a_{ij}^{(c)})$.
5) Generate a new matrix $B = (b_{ij})$ by combining A_r and A_c such that $b_{ij} = \min\{a_{ij}^{(r)}, a_{ij}^{(c)}\}$.
6) Repeat steps $2) - 5)$ to refine the matrix B. The final matrix is the restored image matrix. (Optional step)

We name this procedure as *Procedure CP*. Here "CP" stands for initials of "change-point".

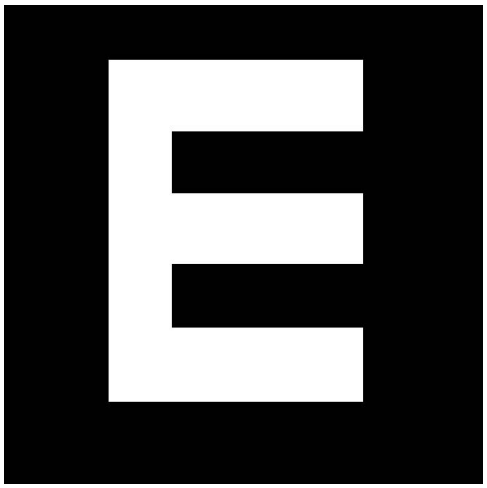


Figure 3. The image of the letter 'E'.

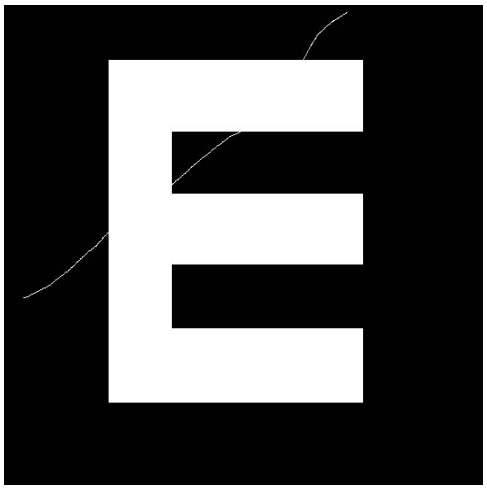


Figure 4. The image of the letter 'E' with some scratches.

IV. TWO EXAMPLES

In this section, we evaluate *Procedure CP* via two examples. We first focus on the example given in [15]. Consider Figures 3-5 below. The image of the letter 'E' is shown in Figure 3. We then add a scratche to the image of the letter 'E'. The resulting image is displayed in Figure 4. In order to examine if our procedure can restore the original image of the letter 'E' under even worse conditions, we contaminate the image displayed in Figure 4 by adding some noise. The resulting image is shown in Figure 5.

To proceed, we first convert the noised image of the letter 'E' with some scratches to the image matrix of dimensions 542×719. We then apply *Procedure CP* to this image matrix for image retrieval. The restored image is displayed in Figure 6, which shows that the image of the letter 'E' is retrieved successfully.

As a second example, we construct an image displayed in Figure 7, which apparently is more complex than the image of the letter 'E' as it is the combination of circle and triangle. Similar to the previous example, we add some scratches to

 ISBN: 978-1-61208-358-2

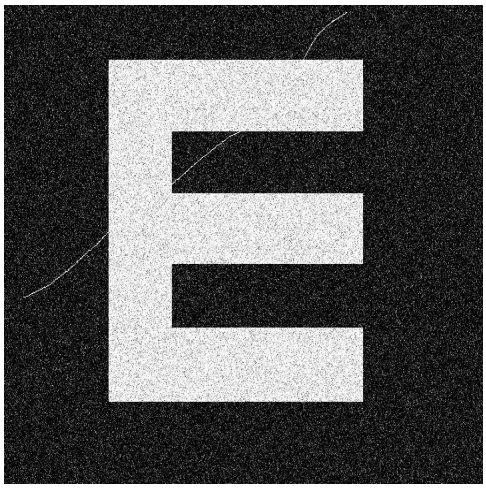


Figure 5. The noised image of the letter 'E' with some added scratches.

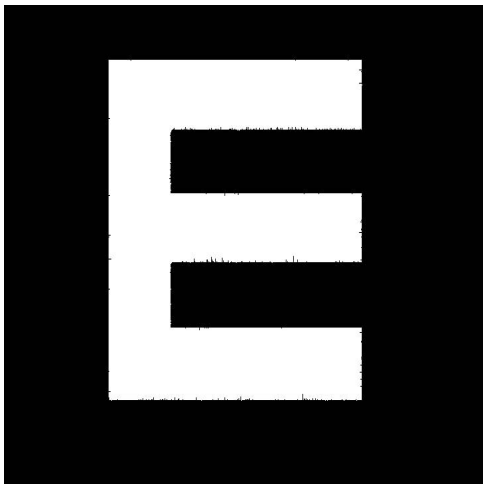


Figure 6. The restored image of the noised image of the letter 'E' with some scratches by applying *Procedure CP*.

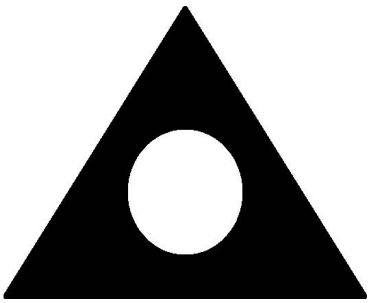


Figure 7. The original image of the second example.

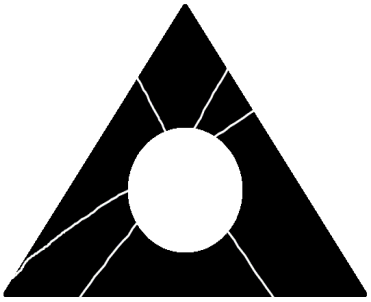


Figure 8. The above image with some scratches.

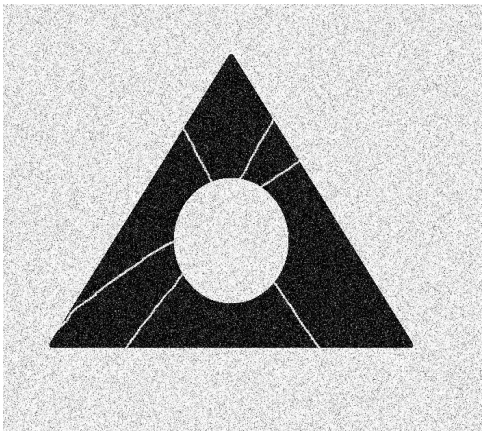


Figure 9. The noised image of the image displayed in Figure 8.

this image (see Figure 8), and further contaminated it by adding some noise (see Figure 9). As above, we first convert the image displayed in Figure 9 to the image matrix, which has dimensions 544×700. We then apply *Procedure CP* to this image matrix for image retrieval. The restored image is displayed in Figure 10, which shows satisfactory performance of *Procedure CP*.

V. CONCLUSION

In this paper, we tackle the image retrieval problem from a different angle. By converting the problem into a multiple change-point detection problem, with the help of *Algorithm WSW*, we propose *Procedure CP* for restoring a noise contaminated black white image with scratches. As demonstrated in two examples, the new method has satisfactory performance in terms of the quality of retrieved images.

We remark that the algorithm given in [16] can be replaced by other multiple change-point detection methods. Even though, we only consider image retrieval for black white images, it may be extended to restore a corrupted gray

 ISBN: 978-1-61208-358-2

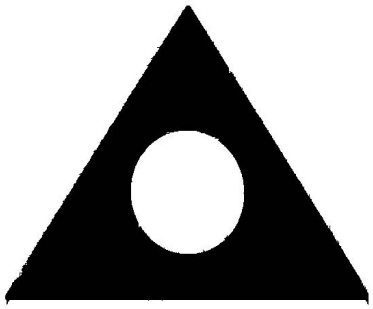


Figure 10. The restored image of the image displayed in Figure 9 by applying *Procedure CP*.

image. The idea used in this paper, i.e., converting the image retrieval problem to a multiple change-point detection problem, may also be applied in other areas such as fast and secure information transmission.

ACKNOWLEDGMENT

This research is partially supported by the Canadian National Science and Engineering Research Council.

REFERENCES

[1] J. Chen and A. K. Gupta, "Parametric Statistical Change Point Analysis. With Applications to Genetics, Medicine, and Finance," 2nd Edition, Boston: Birkhäuser, 2012.

[2] M. Csörgő and L. Horváth, "Limit Theorems in Change-Point Analysis," Chichester: Wiley, 1997.

[3] M. Haidekker, "Advanced Biomedical Image Analysis," Hoboken: Wiley, 2010.

[4] Z. Hlávka, M. Hušková, C. Kirch, and S. G. Meintanis, "Monitoring changes in the error distribution of autoregressive models based on Fourier methods," Test, vol. 21, 2012, pp. 605-634.

[5] M. Hušková and S. G. Meintanis, "Change point analysis based on empirical characteristic function," Metrika, vol. 63, 2006, pp. 145-168.

[6] M. Hušková and S. G. Meintanis, "Change point analysis based on empirical characteristic function of ranks," Sequential Analysis, vol. 25, 2006, pp. 421-436.

[7] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," second ed., Prentice Hall, 2002.

[8] B. K. Gunturk, "Fundamentals of image restoration" in Image Restoration: Fundamentals and Advances, B. K. Gunturk and X. Li, Eds. Boca Raton: CRC Press, 2012, pp. 25-62.

[9] C. Inclán and G. Tiao, G, "Use of cumulative sums of squares for retrospective detection of change of variance," J. Amer. Statist. Assoc., vol. 89, 1994, pp. 913-923.

[10] B. Jähne, "Digital Image Processing," Berlin: Springer, 2005.

[11] X. Li, "Image denoising: Past, present, and future," in Image Restoration: Fundamentals and Advances, B. K. Gunturk and X. Li, Eds. Boca Raton: CRC Press, 2012, pp. 1-24.

[12] E. S. Page, "A test for a change in a parameter occurring at an unknown point," Biometrika, vol. 42, pp. 523527.

[13] G. Qian, X. Shi, and Y. Wu, "A statistical test of change-point in mean that almost surely has zero error probabilities," Aust. N. Z. J. Stat., vol. 55, 2014, pp. 435-454.

[14] P. van Beek, Y. Su, and J. Yang, "Image denoising and restoration based on nonlocal means" in Image Restoration: Fundamentals and Advances, B. K. Gunturk and X. Li, Eds. Boca Raton: CRC Press, 2012, pp. 89-114.

[15] G. Wang and S. Wang, "Recursive computation of Tchebichef moment and its inverse transform," Pattern Recognition, vol.39, 2006, pp. 47-56.

[16] D. Wei, X. Shi, and Y. Wu, "An algorithm for detecting multiple change-points in distribution," preprint.

 ISBN: 978-1-61208-358-2

Co-movement of European Stock Markets based on Association Rule Mining

Youqin Pan, Elizabeth Haran, Saverio Manago
Dept. of Marketing and Decision Science
Salem State University
Salem, USA
Emails :{ypan, eharan, smanago}@salemstate.edu

Yong Hu
Dept. of E-Commerce
Guangdong University of Foreign Studies
Guangdong, China
henryhu200211@163.com

***Abstract*—Due to the fluctuation and complexity of the stock market, it is challenging to capture its non-stationary property and describe its moving tendency. Moreover, globalization increases the interdependence among countries. It is important for investors to understand the co-movement of international stock markets in order to make informed decisions which lead to profit. With the huge amount of data generated by the stock markets, researchers started to explore this problem using different approaches. In this paper, we apply one of the data mining techniques, namely, association rules, to illustrate knowledge patterns and rules of European stock markets. Especially, this paper investigates the co-movement of the European stock market indices with the leading global stock indices. This study shows a strong co-movement between stock market indices of Germany and Unitied Kingdom. Moreover, the European stock markets seem to have strong co-movement with the US stock market. Their co-movement with the Brazil seems to be also strong. However, Brazil stock index does not assume the dominant role, as the US stock index does. This study also shows that there is a weak relationship between European and Japanese stock markets.**

Keywords-co-movement; association rules; stock index; co-integration.

I. INTRODUCTION

International stock market linkages are of great importance for financial decisions of international investors. International diversification reduces total risk of a portfolio. Increase co-movement between asset returns can diminish the advantage of internationally diversified investment portfolios [18]. Changes in co-movement patterns call for an adjustment of portfolios [26].

Forecasting stock index is a challenging task due to its dynamic and complex nature. Forecasting stock index plays an important role in developing effective market trading strategies [14]. Stock markets can be influenced by various factors such as the international environment, government policies, political climate, economic growth, war, and natural disasters. Among these factors, some of them have long-term effect on the markets while others have only short–term effect [28]. Recently, globalization adds more complexity to the movement of stock markets. Globalization in finance and trade increases the interdependence among countries. Such relationships further cause the co-movement of the financial markets between countries. Studies have confirmed that most of the world's stock markets are integrated and associated [22]. Loh [19] claims that understanding the dynamic co-movement between global financial markets plays an important role in predicting stock market returns, allocating assets and diversifying portfolios.

This paper extends the existing literature on stock market co-movement between the European stock markets with that of the US, Brazil, and the Japan. The major European markets include UK, Germany, and Turkey stock markets. The rest of the paper is organized as follows. In Section 2, we give the literature review. Section 3 presents data and research technique. Section 4 presents research findings and discussions. Section 5 concludes the paper.

II. LITERATURE REVIEW

The dynamic interdependence and market integration among major stock exchanges have been investigated by various studies using vector autoregression (VAR) and autoregressive conditional heterosedastic (ARCH) models. Vuran [27] found that the ISE100 index is co-integrated with stock markets of the United Kingdom (FTSE), Brazil (BOVESPA), and Germany (DAX). Floros [9] demonstrated the linkages and co-integration among mature stock indices (such as S&P 500, Nikkei225 and FTSE-100) using a vector error correction model and the Granger-causality approach. Ozdemir and Cakan [23] claimed that there is a strong bidirectional nonlinear causality relationship between the US stock index and the stock market indices of the Japan, France and the UK using non-linear causality tests. Some studies have demonstrated that the U.S stock market has a dominant impact on emerging markets [20] and some developed stock markets such as Japan and France [23]. These studies demonstrated the dynamic causal linkages among international stock market indices.

Contrary to these findings, Chan, Gup and Pan [6] concluded that stock markets are not co-integrated, by analyzing 18 stock market indices. Pascual [24] also found that there is no co-integration relationship between the French, German, and UK stock markets, by using quarterly data. Zhu et al. [30] rejected co-integration relationships between market returns in Shanghai, Shenzhen and Hong Kong. Dimpfl [8] further proved that international financial

 ISBN: 978-1-61208-358-2

markets are not co-integrated in the Engle and Granger [31] sense. In response to such a dilemma, data mining techniques have been introduced to investigate the co-movement between stock markets. Aghabozorgi and The [1] studied stock market co-movement using the three-phase clustering method. Liao and Chou [15] investigated the co-movement of the stock markets of Taiwan and Hong Kong using clustering methods and association rules. Association rules learning discovers interesting correlation patterns among data items in a large dataset by revealing attribute value conditions that co-occur frequently [29]. It aims at uncovering relationship between items that occur together in database. The association rules generated through mining represent an important class of regularities that exist in databases. Nowadays, stock markets across the world have some kinds of connection with each other. In addition, these markets generate huge amount of financial data each day. Thus, mining association rules becomes important since it can provide insights to investors and policy makers to make informed decision.

Since the introduction of the Euro, the European stock markets have become more integrated with the German stock market taking the leadership role [21]. Investing in European stock markets has grown since the European stock markets are one of the most attractive destinations of international funds. The European markets have assumed the leadership role that the US and Japanese markets experienced in the past [3]. Many studies have investigated linkages between European and US markets or among European markets. The majority of the studies focus on the co-movement among developed stock markets [3][21][25]. There are few studies examining the co-movement between developed and emerging markets [20][27]. Berger et al. [5] claimed that emerging markets provide significant diversification potential to investors due to the low integration of these markets with the developed markets worldwide. Therefore, it is important to understand and estimate the co-movement of the emerging markets and the developed markets. Empirical evidence concerning the dynamics of the co-movement of the European markets (including European emerging markets) with the US, Latin America and Japanese stock market is limited. This study intends to fill this gap by investigating the co-movement of the European stock markets within the region and its co-movement with the major global stock markets, such as the USA and Japanese markets.

III. RESEACH METHOD

A. Data

This database was acquired from the UCI Repository of Machine Learning Databases [32]. Data sets include stock indices from both developed and developing markets, as shown in Table 1.

TABLE I. STOCK INDICIES INVESTIGATED IN THIS STUDY

Stock Indices	Detail Information		
	Country	***Stock Index***	***Stock Exchange***
1	US	S&P500	New York Stock Exchange
2	Japan	NIKKEI 225	Tokyo Stock Exchange
3	Germany	DAX	Frankfurt stock Exchange
4	Brazil	BOVESPA	BM&F BOVESPA stock Exchange
5	Turkey	ISE100	Istanbul Stock Exchange
6	United Kingdom	FTSE-100	London Stock Exchange
7	Europe	EU	MSCI European Index
8	Europe	EM	MSCI emerging market index

Among the indices, the MSCI Europe Index captures large and mid-cap representation across 15 developed market countries in Europe, while the MSCI Emerging Markets index captures large and mi-cap representation across 21 emerging markets countries. The entire data set covers the period from January 5, 2009 to February 22, 2011. Missing values were replaced by the previous day's value.

B. Data Mining Techniques

With the development of information technology and software, large amounts of data on stocks traded by the hour and the minute can be easily collected. Therefore, extracting useful information from the huge amount of stock market data becomes critical to investment in stock markets. Data mining techniques have attracted more attention from investors interested in discovering patterns and predicting changes of the stock markets. For instance, Support Vector Machine (SVM) [11][12][13] and Neural Networks [4][7] [12][17] have been used to improve the predictability of stock prices. Recently, mining association rules [1] from large data repositories have attracted considerable attention in various areas including financial domain. Association rule mining was originally used in marketing to discover association rules about which groups of products are likely to be purchased together. The goal of association rule analysis is to discover interesting association and correlation relationships among large sets of data items. Although there are many applications of association rules in the field of data mining problems, it is not common to estimate the co-movement of the stock markets using association rules. In the literature, only a few studies have utilized the association rules to study the behavior of stock markets. Liao, Ho, and Lin [16] implemented association rules learning on the Taiwan stock market to discover knowledge patterns and stock category association to aid portfolio investments. Na and Sohn [22] predicted the movement direction of the Korea Composite Stock Price Index using association rules.

The problem of mining association rules was first introduced by Agrawal et al. [2]. The association rule is defined as follows: Let $I = \{i_1, i_2, \ldots, i_d\}$ be the item set

 ISBN: 978-1-61208-358-2

and D={ t_1, t_2, …..t_N) be the set of all transactions. A transaction tj is said to contain an item set X if X is a subset of t_j. An association rule is an implication expression of the form X->Y, where X and Y are disjoint item sets. The association rule means that the item set Y is likely to occur whenever the item set X occurs. The strength of an association rule can be measured in terms of support and confidence, which indicate the usefulness and certainty of a rule, respectively [10]. Support is denoted as Sup(X, D), which represents the percentage of transaction D that contains the item set X. The higher the support value, the more important the transaction set D is. Accordingly, the support for the rule X->Y is denoted as Sup(XUY, D), which represents the percentage of transactions in D containing both X and Y item sets. The other measure of rule X->Y, called confidence and denoted as Conf(X->Y), which can be expressed in terms of support such that Conf (X->Y) = Sup (X∩Y)/Sup(X, D). It represents the percentage of transactions in D that contain X and also contain Y. In other words, confidence is an estimate of the conditional probability P (Y|X). For the rules with the same confidence level, the rule with the highest support is preferred. However, both measures may not be sufficient to assess the descriptive power of a rule. For example, rules with high confidence may happen by chance. Therefore, the measure lift is used to assess the reliability of support and confidence, it is defined as: Lift=Conf(X->Y)/Sup(Y) [15]. If the lift value is close to 1, it implies that X and Y are independent and the rule is not useful. If the lift value is higher than 1, it indicates that the occurrence of X provides information about Y.

The goal of association rules discovery is to generate all transaction rules that have a certain level of minimum support and confidence. To obtain a small set of useful association rules from this dataset, we conducted association rule analysis using SAS enterprise miner 12.1[33]. The algorithm used to conduct association analysis is ASSOC procedure implemented in SAS data miner [33]. We also set the minimum support value at 10%, and the minimum confidence value at 70%, which were used in previous studies [16][22]. The maximum number of items in an association is set to 2 and 3.

IV. RESULTS

The extracted association rules ordered by confidence have been summarized in Table 2 and Table 3.

TABLE II. SET OF ASSOCIATION RULES BETWEEN TWO STOCK INDICES

Rule	**Condition**
R1: If BOVESPA is up, then EU is up	Confidence 100%, support 40%, Lift 2.22
R2: IF SP is down, then EU is down	Confidence 100%, support 35%, Lift 1.82
R3: IF EU is up, then SP is up	Confidence 100%, support 45%, lift 1.54
R4: If EU is down, then BOVESPA is down	Confidence 100%, support 55%, Lift 1.67
R5: If BOVESPA is down , then EU is down	Confidence 91.67%, support 55%, Lift 1.67
R6: If ISE is down, then EM is down	Confidence 81.82%, support 45%, lift 1.49
R7: If DAX is down, then EU is down	Confidence 80%, support 40%, lift 1.45
R8: If DAX is up, then FTS is up	Confidence 80%, support 40%, lift 1.45
R9: If DAX is down, then BOVESPA is down	Confidence 80%, support 40%, lift 1.33
R10: IF NIKKEI is up, then BOVESPA is down	Confidence 75%, support 45%,, lift 1.25

A. Association Rules

According to Table 2, Rule 1 and Rule 4 indicate that the Brazil stock index has the highest confidence (100%) and high support with MSCI Europe Index. This reveals that BOVESPA is highly correlated with EU. That is, both indices will have the same trend of variation. Rule 2 and Rule 3 also show strong association between EU and SP. That is, US stock market may still play a dominant role in affecting European stock markets especially when the downward trend is present. Within Europe, the major stock index DAX is the driving force to other stock indices within the region according to Rule 7 and Rule 8. However, there seems to be an inverse relationship between NIKKEI and BOVESPA based on Rule 10. In addition, the lift values for all the rules listed in Table 2 are above 1, which implies that these rules are useful.

Table 3 shows association rules among three stock indices. Rule 1 and Rule 3 demonstrate that DAX and FTS are highly correlated with EU, since EU represents the developed markets in Europe. According to Rule 5 and Rule 6, when ISE and DAX are up, EU is up. However, when ISE and DAX are down, EU is also down. As to the global markets, SP still has a strong influence over European stock markets, as shown in Rule 7 and Rule 9.

 ISBN: 978-1-61208-358-2

TABLE III. SET OF ASSOCIATION RULES AMONG THREE STOCK INDICES

Rule	Condition
R1:If FTS and DAX are down, then EU is down	Confidence 97.77%, support 41.01%, Lift 1.86
R2:If EM and DAX are down, then EU is down	Confidence 97.42%, support 35.39%, Lift 1.86
R3:If FTS and DAX are up , then EU is up	Confidence 97.18%, support 38.76%, lift 2.04
R4: If EM and DAX are up, then EU is up	Confidence 96.36%, support 29.78%, Lift 2.03
R5:If ISE and DAX are down, then EU is down	Confidence 96.20%, support 33.15%, Lift 1.83
R6:If ISE and DAX are up, then EU is up	Confidence 95.83%, support 30.15%, lift 2.01
R7:If SP and DAX are down , then EU is down	Confidence 95.43%, support 35.21%, lift 1.82
R8:If NIKKEI is down and EU is up , then DAX is up	Confidence 93.85%, support 22.85%, lift 1.91
R9:If SP and EU are down, then DAX is down	Confidence 92.61%, support 35.21, lift 1.82
R10:If NIKKIE and EU are up, then FTS is up	Confidence 91.94%, support 21.33%,, lift 1.88

B. *Link Maps*

In this study, we also conducted link analysis, which aims at highlighting the linkages between items of interest. The graph link shows the nodes of items within the dataset that are connected to each other. A link presents a connection between two items in a rule. The size of the node is determined by the transaction count, while the weight of the link is related to the confidence of the rule. The higher the confidence is, the thicker the link between nodes. The link graphs are given in Figure 1 and Figure 2. The link graphs are meaningful and interesting. More importantly, it is easy for investors and traders to identify the patterns of movement of these stock indices. Figure 1 shows a strong co-movement among DAX, FTS, EU, SP, and BOVESPA. That is, when these stock indices are rising, they also drive other stock indices (such emerging stock indices) to go up. On the other hand, when these stock indices are down, other stock indices also go down. Therefore, these stocks may be considered as the driving force to the others to move up or move down.

Figure 2 shows that when DAX, FTS and EU indices are down, all other markets tend to be down. The thick links originated from the nodes of these three indices represent that these indices have strong co-movements with other European stock indices and with the global stock indices, such as SP, NIKKEI, and BOVESPA.

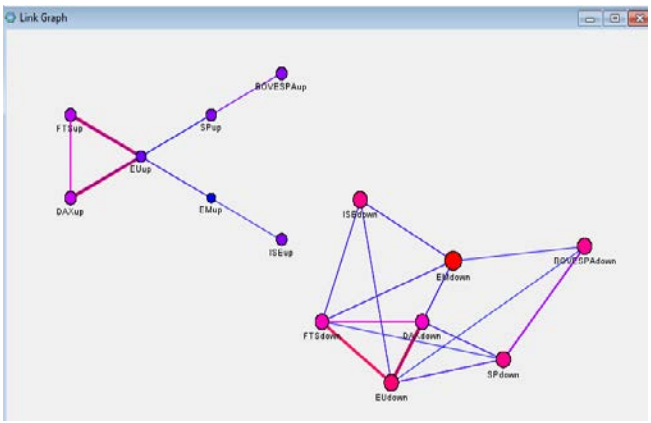


Figure 1. Link Map between two stock indices.

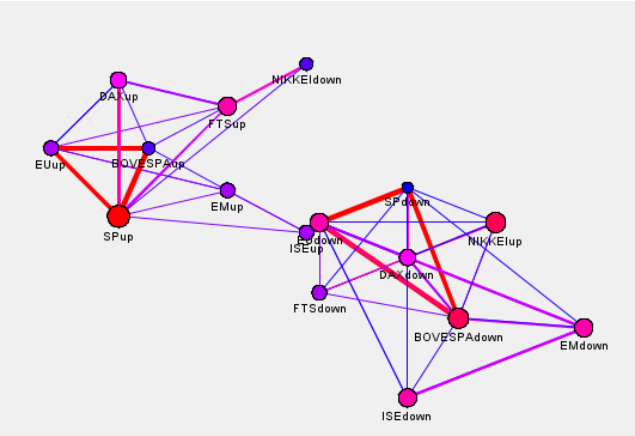


Figure 2. Link Map among three stock indices.

V. CONCLUSION AND FUTURE WORK

In conclusion, we found several interesting associations between European and leading global stock markets. The findings show strong co-movements among European stock indices. The European stock markets seem to have strong co-movement with the US stock market as well. More interestingly, European stock indices also have strong associations with the Brazil index. However, The Brazil stock index does not assume the dominant role as the US stock index does. This study shows that association rules seem to be an appropriate technique for effective exploration of underlying patterns in huge amount of stock market data. We expect that association rules can be used to provide information that can facilitate decision making with regard to predicting stock market returns, allocating assets and diversifying portfolios. However, the results of association analysis need to be interpreted with caution since an association rule does not necessarily imply causality.

 ISBN: 978-1-61208-358-2

REFERENCES

[1] S. Aghabozorgi and Y. H. The, "Stock market co-movement assessment using a three-phase clustering," Expert Systems with Applications, vol. 41, 2014, pp. 1301–1314.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", SIGMOD, vol. 22, 1993, pp. 207–216.

[3] A. Antoniou, G. Pescetto, and A. Violaris, "Modeling International Price Relationships and Interdependencies Between the Stock Index Future Markets of Three EU Countries: A Multivariate Analysis," Journal of Business and Accounting , vol. 30, 2003, pp. 645–667.

[4] G. Armano, M. Marchesi, and A. Murru, "A hybrid genetic-neural architecture for stock indexes forecasting," Information Sciences, vol. 170, 2005, pp. 3-33.

[5] D. Berger, K. Pukthuanthong,. and J. J. Yang, "International diversification with frontier markets," Journal of Financial Economics, vol. 101, 2011, pp. 227-242.

[6] K. C. Chan, B. E., Gup, and M. S. Pan, "International stock market efficiency and integration: A study of eighteen nations," Journal of Business Finance and Accounting, vol. 24, 1997, pp. 803-813.

[7] A. I. Diler, "Predicting direction of ISE national-100 index with back propagation trained neural network," Journal of Istanbul Stock Exchange, vol. 7, 2003, pp. 65-81.

[8] T. Dimpfl, "A note on cointegration of international stock market indices," International Review of Financial Analysis, in press.

[9] C. Floros, "Price linkages between the US, Japan and UK stock markets," Financial Markets and Portfolio Management, vol. 19, 2005, pp. 169-178.

[10] T. Hastie, R. Tibshirani, and J. H. Friedman, "Elements of Statistical Learning: Data Mining, Inference and Prediction," Second Edition, vol. 2, New York : Springer, 2009.

[11] W. Huang, Y., Nakamori, and S.Y. Wang, "Forecasting stock market movement direction with support vector machine," Computers and Operations Research, vol. 32, Oct. 2005, pp. 2513–2522.

[12] Y. Kara , M.A. Boyacioglu, and O. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," Expert Systems with Applications, vol. 38, pp. 5311-5319.

[13] K. J. Kim, " Financial time series forecasting using support vector machines, " Neurocomputing, vol. 55, 2003, pp. 307-319.

[14] M.T. Leung, H., Daouk, and A.S. Chen, " Forecasting stock indices: A comparison of classifcation and level estimation models," International Journal of Forecassting, vol. 16, 2000, pp. 173-190.

[15] S. Liao and S. Chou, "Data Mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio," Expert Systems with Applications, vol. 40, 2013, pp. 1542-1554.

[16] S. Liao, H. Ho, and H. Lin , " Mining stock category association and cluster on Taiwan stock market," Expert Systems with Applications, vol. 35, 2008, pp. 19-29.

[17] Z. Liao and J. Wang, "Forecasting model of global stock index by stochastic time effective neural network," Expert Systems with Applications, vol. 37, 2010, pp. 834–841.

[18] X. Ling and G. Dhesi , " Volatility Spillover and Time-varying conditional correlating between the European and US stock Markets," Global Economy and Finance Journal, vol. 3, 2010, pp. 148-164.

[19] L. Loh, "Co-movement of Asia-Pacific with European and US stock market returns: A cross-time-frequency analysis," Research in International Business and Finance, vol. 29, 2013, pp. 1-13.

[20] A. Masih and R. Masih, "Long and short term dynamic causal transmission amongst international stock markets," Journal of International Money and Finance, vol. 20, 2001, pp. 563-587.

[21] M. Melle , " The EURO Effect on the Integration of European Stock Market,"2003, working paper, Universidad Complutense de Madrid.

[22] S. Na and S. Sohn, " Forecasting changes in Korea Composite Stock Price Index (KOSPI) using association rules," Expert Systems with Applications, vol. 38, 2011, pp. 9046-9049.

[23] Z. A. Ozdemir and E. Cakan , "Non-linear dynamic linkages in the international stock markets, Physica A, vol. 377, 2007, pp. 173–180.

[24] A. G. Pascual, "Assessing European stock markets (co)integration," Economics Letters, vol. 78, 2003, pp. 197–203.

[25] A. Rua and L. C. Nunes, " International comovement of stock market returns: a wavelet analysis," Journal of Empirical Finance, vol.16, 2009, pp. 632–39.

[26] C. Savva and N. Aslanidis, " Stock market integration between new EU member states and the Euro-zone," Empirical Economics, vol. 39, 2010, pp. 337-351.

[27] B. Vuran, " The determination of long-run relationship between ISE100 and international equity indices using cointegration analysis," Istanbul university journal of the school of business administration, vol. 39, 2010, pp. 154-168.

[28] Y. Yang, G. Liu, and Z. Zhang, " Stock market trend prediction based on neural networks, multi-resolution analysis and dynamical reconstruction," Proc. IEEE/IAFE/IAFE/INFORMS 2000 Conference on Computational Intelligence for Financial Engineering, 2000.

[29] D. Zhang and L. Zhou, "Discovering golden nuggets: Data mining in financial application," IEEE Transaction on Systems, Man, and Cybernetics – Part C: Applications and Reviews, vol. 34, 2004, pp. 513–55.

[30] H. Zhu, Z. Lu, and S. Wang, "Causal linkages among Shanghai, Shenzhen, and Hong Kong stock markets," International Journal of Theoretical and Applied Finance, vol. 7, 2004, pp. 135–149

[31] R. F. Engle and C. W. J. Granger, "Co-Integration and Error Correction: Representation, Estimation, and Testing," Econometrica, vol. 55, 1987, pp. 251-276

[32] UCI Machine Learning Repositiory. Retrived January 1, 2013 from: http://archive.ics.uci.edu/ml/

[33] Getting Started with SAS Enterprise Miner 12.1. SAS Institute Inc. Cary, NC, USA

 ISBN: 978-1-61208-358-2

A Method for Measuring Similarity of Simulation Time-Series Data Based on Dynamic Time Warping

Ping Ma, Zhong Zhang, Kaibin Zhao, Yuning Li
Control and Simulation Center
Harbin Institute of Technology
Harbin, P.R. China
pingma@hit.edu.cn, zhangzhong0108@163.com, zhaokaibin1986@163.com, liyuning0043@163.com

***Abstract*—Against the measuring problem of similarity between simulation time-series data and reference time-series data with different lengths or distorted timeline, a similarity measuring method based on DTW is proposed. Considering the trends of time-series data, a new windowing algorithm for Dynamic Time Warping (DTW) is presented. The correctness of the algorithm is proved. On this basis, simulation time-series data and reference time-series data are segmented according to event information, and the overall similarity between them is obtained by calculating DTW distance in different segments. Finally, an example is given to validate the reasonability and validity of the proposed method.**

Keywords-simulation; time-series data; similarity measure; DTW; segmentation

I. INTRODUCTION

As one of the significant means of learning and changing the objective world, the simulation technique is widely used in many fields, such as spaceflight, aviation, electronics and communication [1][2][3]. Simulation system is the application form of simulation technique. Its credibility directly determines the success or failure of the simulation application. Through investigating the difference of simulation system state variables and output variables from actual system state variables and output variables, the simulation credibility can be measured. In other words, we need to investigate whether the similarity between simulation data and reference data is good enough. According to the difference of data arrangement, simulation data can be divided into simulation time-series data and non-time-series data. In the same way, reference data can also be divided to reference time-series data and non-time-series data. This paper studies how to measure the similarity between simulation time-series data and reference time-series data. Without special instructions, simulation data and reference data respectively represent simulation time-series data and reference time-series data in the following content.

Generally speaking, simulation data and reference data are considered a whole separately when investigating the similarity between them, and then their similarity is calculated with an appropriate measuring method. However, various events may occur in the running of actual system, such as "stage-1/stage-2 separation", "fairing jettison", "satellite-rocket separation" during rocket flight. The occurrence of an event usually brings about violent changes in actual system state. Therefore, simulation data and reference data should be divided into segments，and the similarity between two corresponding segments is investigated alone. Combined with all the investigative results, the overall similarity between simulation data and reference data could be obtained. However, the occurrence time of an event is usually advanced or delayed owing to disturbance caused by outside factors. So, the same segments of simulation data and reference data may be distorted on the timeline, which puts forward special requirements on time series similarity measure.

At present, the commonly used time series similarity measures are Euclidean distance [4][5], Edit distance [6] and DTW distance [7][8][9]. The advantages of Euclidean distance include high calculation speed, simple principle and orthogonal transformation invariance. However, it simply regards time series as a point of multidimensional Euclidean space ignoring the time factor of data, and is very sensitive to abnormal points. The most crucial problem is Euclidean distance only applies to measure similarity between time series with synchronization on the timeline. Edit distance is applicable to string sequence. If measuring object is not string sequence, it should be transformed into string sequence by some rule, and then use Edit distance to measure the similarity between two string sequences. But the algorithm applied to transforming non-string sequence into string sequence is not accurate enough, which seriously affect the application effect of Edit distance. Besides, Edit distance also requires two string sequences keep a strict synchronization on the timeline. DTW distance does not require the sampling time of two time series are synchronous, not be sensitive to abnormal points, furthermore, it is able to measure the similarity of time series with different lengths or distorted timeline. Therefore, we introduce DTW into system simulation field, and propose a method for simulation time-series data similarity measure based on DTW. The layout of the paper is as follows: In Section 2, the basic concepts of DTW is introduced. In Section 3, a method for measuring similarity between simulation data and reference data is proposed. In Section 4, the effectiveness of the proposed method is validated through an application example. In Section 5, conclusion is drawn.

II. DTW

DTW is a nonlinear warping technique put forward by Japanese scholar Itakura in the 1960s, and was originally applied in speech recognition. In 1994, DTW was applied to

 ISBN: 978-1-61208-358-2

time-series analysis by Berndt and Clifford [10], and achieved great success. The basic concepts of DTW will be introduced briefly as following.

Definition 1 Suppose time series $X=(x_1,x_2,\ldots,x_m)$, $Y=(y_1, y_2,\ldots,y_n)$, $m\geq 1$, $n\geq 1$. If matrix D satisfies

$$D=\begin{pmatrix} d(x_1,y_n) & d(x_2,y_n) & \cdots & d(x_m,y_n) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_1,y_2) & d(x_2,y_2) & \cdots & d(x_m,y_2) \\ d(x_1,y_1) & d(x_2,y_1) & \cdots & d(x_m,y_1) \end{pmatrix} \quad (1)$$

where $d(x_i, y_j)=|x_i-y_j|$, $i=1,2,\ldots,m$, $j=1,2,\ldots,n$, then we say D is the distance matrix of X and Y.

Definition 2 Suppose time series $X=(x_1,x_2,\ldots,x_m)$, $Y=(y_1, y_2,\ldots,y_n)$, $m\geq 1$, $n\geq 1$. If sequences $P=\langle p_1,p_2,\ldots,p_K\rangle$, $p_k\in\{(a,b)|a=1,2,\ldots,m,\ b=1,2,\ldots,n\}$, $k=1,2,\ldots,K$, $\max(m,n)\leq K\leq m+n+1$, satisfies

(1) Boundary conditions: $p_1=(1,1)$, $p_K=(m,n)$;

(2) Continuity: Given $p_k=(a,b)$, then $p_{k+1}=(a',b')$, where $|a'-a|\leq 1$ and $|b'-b|\leq 1$;

(3) Monotonicity: Given $p_k=(a,b)$, then $p_{k+1}=(a',b')$, where $a'-a\geq 0$ and $b'-b>0$, or $a'-a>0$ and $b'-b\geq 0$, then we say P is a warping path between X and Y.

There are exponentially many warping paths that satisfy the above conditions, as shown in Figure 1. However, we are only interested in the path which has the minimum cumulative distance, and regard it as the basis of measuring similarity degree between two time series.

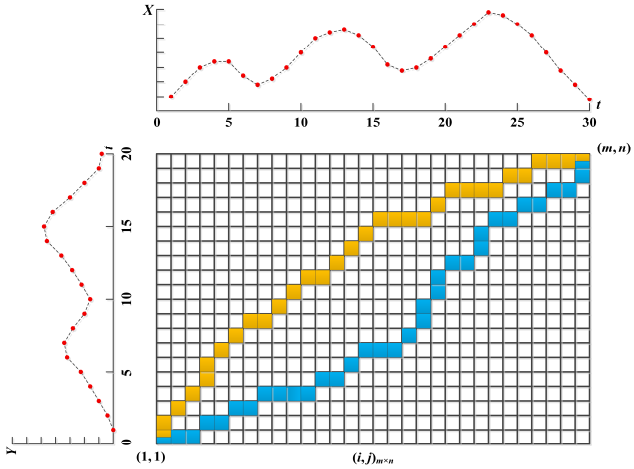


Figure 1. Warping paths between two time series

Definition 3 Suppose time series $X=(x_1,x_2,\ldots,x_m)$, $Y=(y_1, y_2,\ldots,y_n)$, $m\geq 1$, $n\geq 1$, then the DTW distance between X and Y is

$$DTW(X,Y)=\min\left\{\sum_{k=1}^{K} d(x_{a_k},y_{b_k})\,\middle|\,\forall P=\langle(a_1,b_1),\ldots,(a_K,b_K)\rangle\right\} \quad (2)$$

The smaller $DTW(X, Y)$ is, the higher degree of similarity between X and Y is. Otherwise it is lower.

III. SIMILARITY MEASURING METHOD BASED ON DTW

This section discusses the warping window of DTW first, and constructs a new windowing algorithm. On the basis of the new windowing algorithm, a method for measuring similarity between simulation data and reference data is proposed.

A. Warping Window of DTW

In the application of DTW, swing scope of warping path is generally limited. On the one hand, this can narrow the scope to search optimal warping path, and improve calculation speed of DTW distance. On the other hand, even more important, this can avoid a morbid warping path, and ensure the reasonableness of calculation result of DTW distance. At present, the most common windowing algorithm is: for any point $p_k=(a_k, b_k)$ in the warping path, it is required to satisfy $|a_k-b_k|\leq L$. It is obvious that the value of L directly affects the reasonableness of calculation result of DTW distance. However, the value of L is generally given by experts in practice, which has strong subjectivity. In addition, warping window of the same width cannot precisely describe the distortion of time series in timeline.

Against the above problems, a new windowing algorithm called TS algorithm is presented in this section. Its basic idea is through analyzing the trend of time series, data matching is limited to segments in which two time series have the same trend. So, it can effectively prevent the excessive distortion of warping path. The steps of TS algorithm are as follows:

Step 1 Calculate the trend sequence of time series $X=(x_1, x_2,\ldots,x_m)$ and $Y=(y_1,y_2,\ldots,y_n)$, and get $X'=(x_1',x_2',\ldots,x_{m-1}')$, $Y'=(y_1',y_2',\ldots,y_{n-1}')$, where

$$x_i'=\begin{cases} 0, & x_i=x_{i+1}; \\ 2(x_{i+1}-x_i)/(|x_i|+|x_{i+1}|), & \text{otherwise}; \end{cases} \quad (3)$$

$$y_j'=\begin{cases} 0, & y_j=y_{j+1}; \\ 2(y_{j+1}-y_j)/(|y_j|+|y_{j+1}|), & \text{otherwise}; \end{cases} \quad (4)$$

$i=1,2,\ldots,m-1$, $j=1,2,\ldots,n-1$.

Step 2 Calculate the DTW distance between X' and Y', and get optimal warping path $P'=\langle p_1',p_2',\ldots,p_K'\rangle$;

Step 3 On the basis of warping path P', we get warping window $W=\{(a,b)|a=1,2,\ldots,m,\ b=1,2,\ldots,n,\{(a-1,b-1),(a-1,b),(a,b-1),(a,b)\}\cap\{p_1',p_2',\ldots,p_K'\}\neq\varnothing\}$, and then calculate the DTW distance between X and Y.

TS algorithm has great objectiveness due to no need of expert knowledge, and can precisely describe the distortion of time series in timeline.

Theorem 1 Suppose time series $X=(x_1,x_2,\ldots,x_m)$ and $Y=(y_1,y_2,\ldots,y_n)$. W is the warping window calculated with TS algorithm, then there exists at least one warping path between X and Y in W.

Proof For proving that there exists a warping path between X and Y in W, we only need to prove the following two conditions:

(1) $(1,1)\in W$, $(m,n)\in W$;

 ISBN: 978-1-61208-358-2

Suppose P' is optimal warping path between X' and Y', so, $p'_1=(1,1)$ and $p'_K=(m-1,n-1)$, and then

$$\{(0,0),(0,1),(1,0),(1,1)\}\cap\{p'_1,p'_2,\ldots,p'_K\}=(1,1)\neq\varnothing$$

$$\{(m-1,n-1),(m-1,n),(m,n-1),(m,m)\}\cap\{p'_1,p'_2,\ldots,p'_K\}=(m-1,n-1)\neq\varnothing$$

Therefore, $(1,1)\in W$, $(m,n)\in W$.

(2) If $(a,b)\in W$ and $(a,b)\neq(m,n)$, then $\{(a,b+1),(a+1,b),(a+1,b+1)\}\in W\neq\varnothing$;

Suppose P' is optimal warping path between X' and Y', so for $\forall(a',b')\in\{p'_1,p'_2,\ldots,p'_K\}$, $(a',b')\neq(m-1,n-1)$, we all have

$$\{(a',b'+1),(a'+1,b'),(a'+1,b'+1)\}\cap\{p'_1,p'_2,\ldots,p'_K\}\neq\varnothing$$

Then

$$\left.\begin{array}{l}(a,b)\in W\\(a,b)\neq(m,n)\end{array}\right\}\Rightarrow\{(a-1,b-1),(a-1,b),(a,b-1),(a,b)\}\cap\{p'_1,p'_2,\ldots,p'_K\}\neq\varnothing$$

$$\Rightarrow\{\{(a-1,b),(a,b-1),(a,b)\}\cup\{(a-1,b+1),(a,b),(a,b+1)\}\cup\{(a,b),(a+1,b-1),(a+1,b)\}\cup\{(a,b+1),(a+1,b),(a+1,b+1)\}\}\cap\{p'_1,p'_2,\ldots,p'_K\}\neq\varnothing$$

$$\Rightarrow\{\{(a-1,b),(a-1,b+1),(a,b),(a,b+1)\}\cap\{p'_1,p'_2,\ldots,p'_K\}\}\cup\{\{(a,b-1),(a,b),(a+1,b-1),(a+1,b)\}\cap\{p'_1,p'_2,\ldots,p'_K\}\}\cup\{\{(a,b),(a,b+1),(a+1,b),(a+1,b+1)\}\cap\{p'_1,p'_2,\ldots,p'_K\}\}\neq\varnothing$$

$$\Rightarrow\{(a,b+1),(a+1,b),(a+1,b+1)\}\cap W\neq\varnothing$$

From the above, we know that there exists at least one warping path between X and Y in W.

B. Steps of Algorithm

Suppose reference data $R=(r(t_1),r(t_2),\ldots,r(t_n))$, simulation data $S=(s(t_1),s(t_2),\ldots,s(t_n))$, $t_1<t_2<\ldots<t_n$. Event sequence $E=(e_1,e_2,\ldots,e_K)$ is happened to actual system during $t_1\sim t_n$, and its happening time is $T_1=(t_{1,1},t_{1,2},\ldots,t_{1,K})$; The same event sequence $E=(e_1,e_2,\ldots,e_K)$ is happened to simulation system during $t_1\sim t_n$, and its happening time is $T_2=(t_{2,1},t_{2,2},\ldots,t_{2,K})$.

To measure the similarity between reference data R and simulation data S, a similarity measuring method is proposed, in combination with research achievement in the DTW theory. Its steps are as follows:

Step 1 According to the event sequence E, R is segmented into $R=\langle R_1,R_2,\ldots,R_{K+1}\rangle$, and S is segmented into $S=\langle S_1,S_2,\ldots,S_{K+1}\rangle$, as shown in (5) and (6).

Step 2 Using group AHP to determine the weight of each segment. This step includes the following four substeps:

Step 2.1 According to the actual situation of simulation object, a number of experts with relevant background and different knowledge structure are invited to set up an evaluation team $P=\{p_1,p_2,\ldots,p_m\}$. The subjective weight of expert p_i is obtained from professional title, evaluation mode and the degree of familiarity with simulation object. The formula is as follows:

$$\lambda_{p_i}^{(s)}=\frac{l_{i,1}\times l_{i,2}\times l_{i,3}}{\sum_{j=1}^{m}l_{j,1}\times l_{j,2}\times l_{j,3}} \tag{7}$$

where $l_{i,1}$, $l_{i,2}$, $l_{i,3}$ are valued as shown in Table 1;

TABLE I. THE FACTOR VALUE OF EXPERT SUBJECTIVE WEIGHT

Professional title	$l_{i,1}$	Evaluation mode	$l_{i,2}$	The degree of familiarity with simulation object	$l_{i,3}$
professor/researcher/professor senior engineer	4	two-way anonymous	3	very familiar	4
associate professor/associate researcher/ senior engineer	3	one-way anonymous	2	familiar	3
lecturer/assistant researcher/engineer	2	two-way open	1	general understanding	2
ta/intern researcher/assistant engineer	1			not familiar	1

Step 2.2 The difference between R and S in each segment has different degree of influence on simulation credibility. So, expert e_i is invited to make comparison between the influence degrees, then we get judgement matrix

$$A_i=\left(a_{jk}^{(i)}\right)_{(K+1)\times(K+1)} \tag{8}$$

$i=1,2,\ldots,m$;

Step 2.3 The consistency ratio of judgement matrix reflects the thinking logicality of expert when he makes a judgement. The greater the consistency ratio, the more serious the logic conflict of judgement matrix, the less objective weight of relevant expert; On the contrary, the more objective weight of relevant expert. Hence, the objective weight of

$$R_k=\begin{cases}(r(t_1),r(t_2),\ldots,r(t_i)), & k=1,\ t_i\le t_{1,k}<t_{i+1}\\(r(t_i),r(t_{i+1}),\ldots,r(t_{i+j})), & 2\le k\le K,\ t_{i-1}\le t_{1,k-1}<t_i,\ t_{i+j}\le t_{1,k}<t_{i+j+1}\\(r(t_i),r(t_{i+1}),\ldots,r(t_n)), & k=K+1,\ t_{i-1}\le t_{1,k-1}<t_i\end{cases} \tag{5}$$

$$S_k=\begin{cases}(s(t_1),s(t_2),\ldots,s(t_i)), & k=1,\ t_i\le t_{2,k}<t_{i+1}\\(s(t_i),s(t_{i+1}),\ldots,s(t_{i+j})), & 2\le k\le K,\ t_{i-1}\le t_{2,k-1}<t_i,\ t_{i+j}\le t_{2,k}<t_{i+j+1}\\(s(t_i),s(t_{i+1}),\ldots,s(t_n)), & k=K+1,\ t_{i-1}\le t_{2,k-1}<t_i\end{cases} \tag{6}$$

 ISBN: 978-1-61208-358-2

expert p_i is obtained from the consistency ratio. The formula is as follows:

$$\lambda_{p_i}^{(c)} = \frac{1/e^{CR(A_i)}}{\sum_{j=1}^{m}\left(1/e^{CR(A_j)}\right)} \tag{8}$$

where $CR(A_i)$ is the consistency ratio of A_i, whose solving process is referred to literature [11];

Step 2.4 According to expert subjective weight and objective weight, expert synthesis weight is calculated by the following formula:

$$\lambda_{p_i} = \tau\lambda_{p_i}^{(s)} + (1-\tau)\lambda_{p_i}^{(c)} \tag{9}$$

where τ is an adjusting coefficient, $0 \le \tau \le 1$. The smaller τ is, the more attention pay to the actual performance of expert in evaluation work; The greater τ is, the more attention pay to the qualification of expert;

Step 2.5 Combined with expert synthesis weight, synthesis judgement matrix is obtained by using Hadamard convex combination to aggregate individual judgement matrices as follows:

$$\overline{A} = \left(\overline{a}_{jk}\right)_{(K+1)\times(K+1)} = \left(\prod_{i=1}^{m}\left(a_{jk}^{(i)}\right)^{\lambda_{p_i}}\right)_{(K+1)\times(K+1)} \tag{10}$$

At last, the weight of each segment is calculated by $\overline{A}$, then we get $\lambda_1, \lambda_2, \ldots, \lambda_{K+1}$;

Step 3 Calculate the average DTW distance R_i and S_i based on TS algorithm, the formula is as follows:

$$\overline{DTW}(R_i, S_i) = \frac{DTW(R_i, S_i)}{l_i} \tag{11}$$

where $DTW(R_i, S_i)$ is the length of optimal warping path, l_i is the number of steps in optimal warping path, i=1,2,…,K+1;

Step 4 Calculate the similarity between R and S, the formula is as follows:

$$\Psi(R,S) = \frac{1}{1+\delta} \tag{12}$$

where $\delta = \dfrac{\sum_{i=1}^{K+1}\lambda_i \cdot \overline{DTW}(R_i, S_i)}{\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}\min\left(r^2(t_i), s^2(t_i)\right)}}$, $0 \le \Psi(R,S) \le 1$.

IV. APPLICATION EXAMPLE

To avoid interception by our anti-aircraft missile, enemy attack missiles always tend to be changing orbit dramatically. Guidance law needs to be designed for our missile to intercept enemy missile effectively. A guidance simulation model is set up to validate guidance law. Take lateral acceleration output of this model for example, the effectiveness of the proposed method in this paper is validated.

Assume the motion of the target is as follows:

1) Before orbital maneuver: keep 8000m height, do horizontal straight line motion with a speed in 600m/s;

2) After orbital maneuver: keep 8000m height, do sine motion around the straight line, the sine law: 150sin(0.3t)m;

3) The maneuver time t(s) obeys the normal distribution: $N(6,0.2^2)$.

Data sampling interval: 0.02s, Sampling period: 0s~10s, 500 samples in total, as shown in Fig.2, where R is reference data of lateral acceleration and S is simulation data of lateral acceleration.

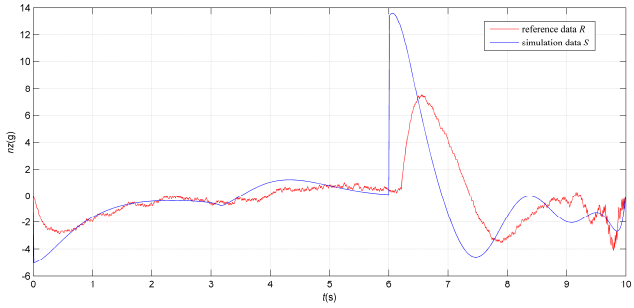


Figure 2. Comparison between reference and simulation lateral acceleration

Using the proposed method to calculate similarity between R and S as follows:

First, according to maneuver time of target, R is segmented into $R=\langle R_1, R_2\rangle$, and S is segmented into $S=\langle S_1, S_2\rangle$, where $R_1=(r(t_1),r(t_2),\ldots,r(t_{310}))$, $R_2=(r(t_{310}),r(t_{311}),\ldots,r(t_{500}))$, $S_1=(s(t_1),s(t_2),\ldots,s(t_{300}))$, $S_2=(s(t_{300}),s(t_{301}),\ldots,s(t_{500}))$.

Then, invite three experts to constitute an evaluation team $P=\{p_1,p_2,p_3\}$, and use (7) to calculate subjective weight of expert as follows:

$$\lambda_{p_1}^{(s)} = \frac{4\times3\times4}{4\times3\times4+4\times3\times2+3\times3\times3} = 0.485$$

$\lambda_{p_2}^{(s)} = 0.242$, $\lambda_{p_3}^{(s)} = 0.273$.

The difference between R and S has influence on simulation credibility, and the influence degree is different before and after target orbital maneuver. In the light of this, each expert gives an individual judgement matrix as follows:

$$A_1 = \begin{pmatrix} 1 & 1/3 \\ 3 & 1 \end{pmatrix},\ A_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\ A_3 = \begin{pmatrix} 1 & 1/4 \\ 4 & 1 \end{pmatrix}$$

In combination with A_1, A_2, and A_3, use (9) to calculate objective weight of expert: $\lambda_{p_1}^{(c)} = \lambda_{p_2}^{(c)} = \lambda_{p_3}^{(c)} = 0.333$; On this basis, use (10) to calculate synthesis weight of expert: $\lambda_{p_1} = \tau\lambda_{p_1}^{(s)} + (1-\tau)\lambda_{p_1}^{(c)} = 0.447$, $\lambda_{p_2} = 0.265$, $\lambda_{p_3} = 0.288$, where $\tau = 0.75$.

In combination with synthesis weight of expert, use (11) to calculate comprehensive judgement matrix which is

$$\overline{A} = \begin{pmatrix} 1 & 0.411 \\ 2.436 & 1 \end{pmatrix}$$

And then we get $\lambda_1 = 0.291$, $\lambda_2 = 0.709$.

Next, the average DTW distances between R_1 and S_1, R_2 and S_2 are calculated respectively based on TS algorithm: $\overline{DTW}(R_1, S_1) = 0.572$, $\overline{DTW}(R_2, S_2) = 1.437$, their waring paths are shown in Fig.3.

Finally, calculate the overall similarity between R and S

 ISBN: 978-1-61208-358-2

$$\Psi(R,S)=\frac{1}{1+\dfrac{0.291\times0.572+0.709\times1.437}{1.655}}=0.583$$

As contrast, use Euclidean distance to calculate the similarity between R and S, and get $\Psi'(R,S)=0.380$. In calculation process of $\Psi(R,S)$ and $\Psi'(R,S)$, the alignment way of time series is shown in Fig.4.

It is clearly observed that the proposed method can measure similarity between simulation data and reference data with different length, and it has a strong robustness on the distortion of timeline. $\Psi(R,S)>\Psi'(R,S)$ also shows that the proposed method make up for the deficiency of traditional simulation data similarity measures, and the measurement result is more objective and reasonable.

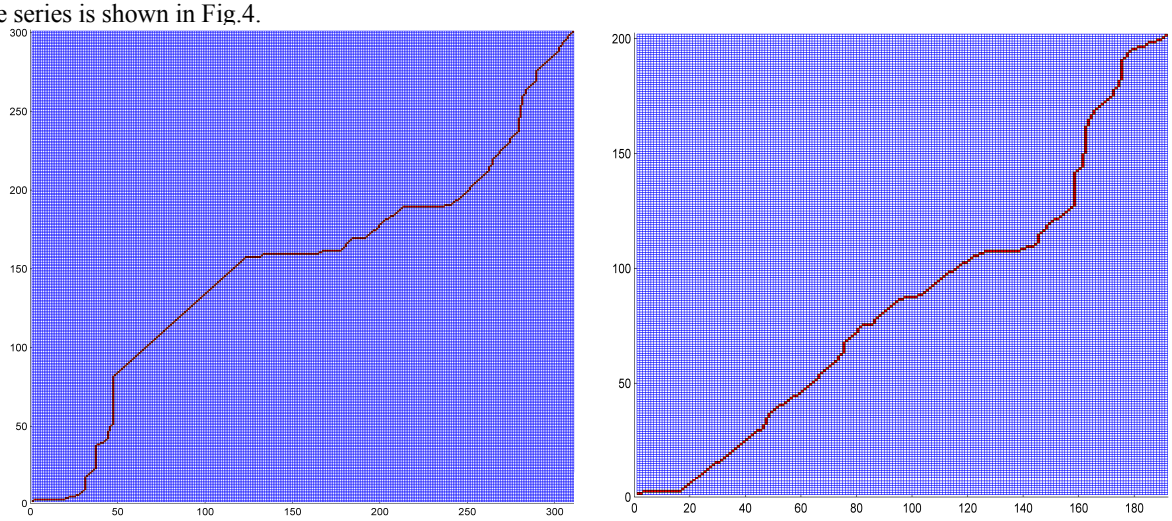
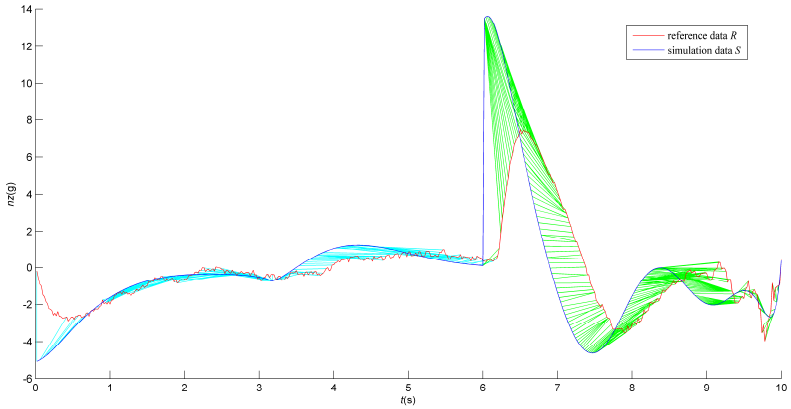
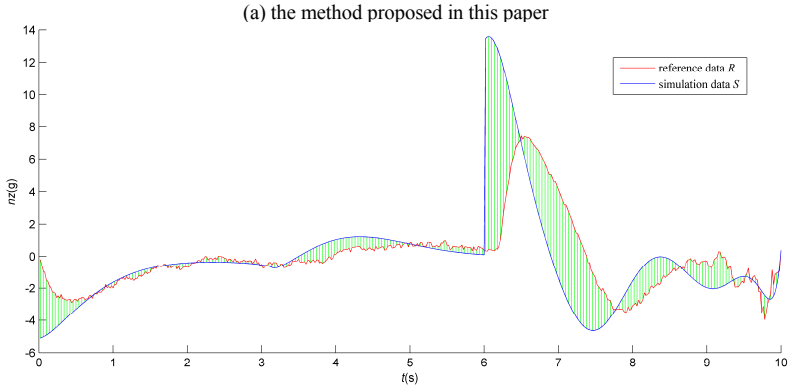


Figure 3. Warping paths of DTW distance



(a) the method proposed in this paper



(b) Euclidean distance

Figure 4. The alignment way of time series

V. CONCLUSION

The similarity between simulation data and reference data is generally investigated when we evaluate the credibility of simulation system. As system states usually change sharply after the occurrence of some event, so, data need to be segmented for analysis. Against simulation time and actual time are not in synch, a similarity measuring method based on DTW is proposed, using a new windowing algorithm for DTW. Application example shows that the proposed method can effectively measure the similarity between simulation data and reference data with different lengths or distorted timeline, and has a good prospect in the field of simulation model validation.

ACKNOWLEDGMENT

This research is supported by the Innovative Team Program of the National Natural Science Foundation of China under Grant No.61021002.

REFERENCES

[1] B. Gyss, "Physics-Based Human Biomechanical Simulation for Long-Duration Space-Flight-Related Applications," Journal of the British Interplanetary Society, vol. 66, Sep. 2013, pp. 275-277.

[2] I. Stevanovic, S. Skibin, M. Masti and M. Laitinen, "Behavioral Modeling of Chokes for EMI Simulations in Power Electronics," IEEE Transactions on Power Electronics, vol. 28, Feb. 2013, pp. 695-705.

[3] Z. Z. Wang, F. C. Jia, E. R. Galea and J. Ewer, "Computational Fluid Dynamics Simulation of a Post-Crash Aircraft Fire Test," Journal of Aircraft, vol. 50, Jan. 2013, pp. 164-175.

[4] C. J. Liu, "Discriminant Analysis and Similarity Measure," Pattern Recognition, vol. 47, Jan. 2014, pp. 359-367.

[5] T. Mcloughlin, M. W. Jones, R. S. Laramme, R. Malki, I. Masters, et al., "Similarity Measures for Enhancing Interactive Streamline Seeding," IEEE Transactions on Visualization and Computer Graphics, vol. 19, Aug. 2013, pp. 1342-1353.

[6] E. S. Ristad and P. N. Yianilos, "Learning String-Edit Distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, May. 1998, pp. 522-532.

[7] E. J. Keogh and M. J. Pazzani, "Derivative Dynamic Time Warping," Proceedings of the 2001 SIAM International Conference on Data Mining, Apr. 2001, pp. 1-11.

[8] Z. Banko and J. Abonyi, "Correlation Based Dynamic Time Warping of Multivariate Time Series," Expert Systems with Applications, vol. 39, Dec. 2012, pp.12814-12823.

[9] H. Sarin, M. Kokkolaras, G. Hulbert and et al., "A Comprehensive Metric for Comparing Time Histories in Validation of Simulation Models with Empahais on Vehicle Safety Applications," Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, August. 2008, pp. 1275-1286.

[10] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Data-bases, Jul. 1994, pp. 359-370.

[11] J. M. Moreno-Jimenez, J. Aguaron, and M. T. Escobar, "The Core of Consistency in AHP-Group Decision Making," Group Decision and Negotiation, vol. 17, May. 2008, pp. 249-265.

 ISBN: 978-1-61208-358-2

Validation of Simulation Model Based on Combined Consistency Analysis of Data and Feature

Ming Yang, Wei Li, Lingyun Lu, Song Jiao
Control and Simulation Center
Harbin Institute of Technology
Harbin, P. R. China
e-mail: myang@hit.edu.cn, fleehit@163.com, lulingyun987@163.com, jiaosong1985@163.com

***Abstract*—Comparing the simulation output with reference output is an important measure to validate a simulation model. The classic methods either analyze the data consistency of simulation output and reference output or the consistency of their features. Considering the two types of consistencies simultaneously, a validation method of simulation model based on combined consistency analysis of data and feature is proposed. The measurement model of data consistency integrating the proximity of spatial position and similarity of geometrical shape is presented, and the measurement model of feature consistency based on relative error is given. Besides, the weight of each consistency index was determined by analyzing the correlation among the consistency indexes, and the consistency integration model is given. Finally, the validation method is applied to validate a simulation model of servo-control system.**

Keywords-model validation; data consistency; feature consistency; correlation analysis.

I. INTRODUCTION

Simulation has been widely used in many fields, such as military, economic and social areas, with the advantages of economy, security and efficiency. Because simulation is a scientific research based on models, the credibility of model attracts much attention. The simulation model validation is the primary means to research the credibility of model, and the consistency analysis of simulation output and the reference output is an important method for simulation model validation [1][2].

As early as 1962, Biggs made an assessment for the credibility of the "dogs" missile model [3]. Fishman and Kiviat used spectral analysis to assess the credibility of the queuing model [4]. Hermann made a consistency validation between the real system and simulation model, based on the intrinsic features or typical events [5]. Mckenny proposed some methods based on the analysis of variance, Kolmogorov-Smirnov test and χ^2 test [6]. Kheir and Holmes utilized Theil inequality coefficient (TIC) method to validate the effectiveness of a missile simulation system [7]. Through the analysis of features of grey correlation model, Wei and Li made an application to assess the credibility of a missile simulation system [8]. Damborg assessed the credibility of econometric models with the error analysis methods [9]. Liu, Liu and Zhang discussed the relation between credibility and similarity, and gave a quantitative method of simulation model credibility according to the similarity [10]. Moreover, Balci, Sargent and Kleijnen gave some summaries to the validation of simulation model [11][12][13].

The analysis of classical methods above shows that, there are mainly two research ideas, namely, 1) the features we concerned were extracted from the output data of simulation and reference, such as the rise time, overshoot and steady-state error in controlling response. Then, feature consistency can be analyzed via variance analysis and hypothesis test, etc., and 2) the data consistency of simulation output and reference output was analyzed directly, such as TIC method and gray correlation method, etc. The data consistency of simulation output and reference output reflects their panoramic consistency, but ignores the consistencies of some detailed features easily. For some simulation applications, the data consistency and feature consistency are focused on. So, the two types of consistencies should be considered simultaneously in the validation of simulation model.

To solve the above problem in terms of continuous dynamic simulation of multi-output, this paper presents a new method of simulation model validation, considering multiply features. The advantages of this new method are more comprehensive, convincing and reliable. The paper is organized as follows: In Section 2, the research problems will be described and analyzed. In Section 3, the measurement models of data consistency and feature consistency will be given. In Section 4, the integrated model of consistency will be given. In Section 5, the effectiveness of this method will be shown through application examples. In the last Section, summary is drawn and future research is discussed.

II. PROBLEM DESCRIPTION AND ANALYSIS

Assuming that $U_s = [u_{s1}, u_{s2}, \cdots, u_{sk}]^T$ and $U_r = [u_{r1}, u_{r2}, \cdots, u_{rk}]^T$ denote the inputs of simulation model and reference system respectively. Their outputs are represented as $Y_s = [y_{s1}, y_{s2}, \cdots, y_{sm}]^T$ and $Y_r = [y_{r1}, y_{r2}, \cdots, y_{rm}]^T$. The simulation model researched in the paper is described as follows:

$$\begin{cases} \dot{X}_s = F(X_s, U_s, T) \\ Y_s = G(X_s, U_s, T) \end{cases} \quad (1)$$

 ISBN: 978-1-61208-358-2

$$F(X_s,U_s,T)=\begin{bmatrix} f_1(X_s,U_s,T) \\ \vdots \\ f_n(X_s,U_s,T) \end{bmatrix} \quad (2)$$

$$G(X_s,U_s,T)=\begin{bmatrix} g_1(X_s,U_s,T) \\ \vdots \\ g_m(X_s,U_s,T) \end{bmatrix} \quad (3)$$

where $X_s=[x_{s1},x_{s2},\cdots,x_{sn}]^T$ is state variable, $F(X_s,U_s,T)$ is state equation and $G(X_s,U_s,T)$ is output equation, T is time series.

Assuming that $C(Y_s,Y_r)$ denotes the consistency between Y_s and Y_r when $U_s=U_r$ (the consistency of simulation output for short), and $C(Y_s,Y_r)\in(0,1]$. When $C(Y_s,Y_r)=1$, it indicates that Y_s is the same with Y_r exactly, i.e., the simulation model is quite credible. If the consistency between Y_s and Y_r becomes more and more bad, i.e., the simulation model is more and more incredible, then $C(Y_s,Y_r)\to 0$.

Assuming that $D(y_{si},y_{ri})$, $i=1,2,\cdots,m$ denotes the data consistency between y_{si} and y_{ri} when $U_s=U_r$ (the consistency of y_{si} for short), and $D(y_{si},y_{ri})\in(0,1]$. When $D(y_{si},y_{ri})=1$, it indicates that y_{si} is the same with y_{ri} exactly. If the data consistency between y_{si} and y_{ri} becomes more and more bad, then $D(y_{si},y_{ri})\to 0$.

Assuming that $d(y_{si},y_{ri})$ and $s(y_{si},y_{ri})$ denote "proximity" of position (i.e., proximity of spatial position) and "similarity" of shape (i.e., similarity of geometrical shape) between y_{si} and y_{ri} when $U_s=U_r$. The definition is as follows:

$$D(y_{si},y_{ri})=G(d(y_{si},y_{ri}),s(y_{si},y_{ri})),\ i=1,2,\cdots,m \quad (4)$$

where $G(\bullet)$ is the integrated model of $d(y_{si},y_{ri})$ and $s(y_{si},y_{ri})$.

Assuming that $C_s=[c_{s1},c_{s2},\cdots,c_{sl}]^T$ and $C_r=[c_{r1},c_{r2},\cdots,c_{rl}]^T$ denote features extracted from Y_s and Y_r respectively. A definition is as follows:

$$C_s=H(Y_s),\ C_r=H(Y_r) \quad (5)$$

where $H(\bullet)$ is the feature extraction model.

Assuming that $V(c_{sj},c_{rj})$, $j=1,2,\cdots,l$ denotes the feature consistency between c_{sj} and c_{rj} when $U_s=U_r$, and $V(c_{sj},c_{rj})\in(0,1]$. When $V(c_{sj},c_{rj})=1$, it indicates that c_{sj} is the same with c_{rj} exactly. If the feature consistency between c_{sj} and c_{rj} becomes more and more bad, then $V(c_{sj},c_{rj})\to 0$.

To gain $C(Y_s,Y_r)$, the classic methods only consider $D(y_{si},y_{ri})$, $i=1,2,\cdots,m$ or $V(c_{sj},c_{rj})$, $j=1,2,\cdots,l$. For some simulation application, the two types of consistencies should be considered simultaneously. As shown in Fig. 1, the index system of simulation output consistency is given.

From the above, we can notice that the application of method proposed in this paper has two difficulties: 1) how to measure the data consistency and feature consistency; 2) how to deal with the correlation among the consistencies of data and features and integrate them to gain the consistencies of simulation output.

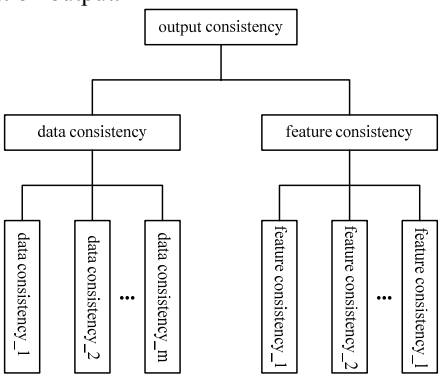


Figure 1. Index system of simulation output consistency

III. CONSISTENCY MEASUREMENT MODEL

In this section, two types of consistency measurement model are presented.

A. Data consistency measurement model

The reference output and the simulation output of a component are denoted by $y_r=<y_r(1),y_r(2),\cdots y_r(p)>$ and $y_s=<y_s(1),y_s(2),\cdots y_s(p)>$, respectively. The data consistency can be obtained by considering the "proximity" of position and the "similarity" of shape comprehensively.

The simulation model is validated by the difference of position between y_r and y_s, which is depicted by the coefficient of TIC in [7]. The formula is as follows:

$$U_{TIC}(y_r,y_s)=\frac{\sqrt{\frac{1}{p}\sum_{i=1}^{p}(y_r(i)-y_s(i))^2}}{\sqrt{\frac{1}{p}\sum_{i=1}^{p}y_r(i)^2}+\sqrt{\frac{1}{p}\sum_{i=1}^{p}y_s(i)^2}} \quad (6)$$

where $U_{TIC}(y_r,y_s)$ is the coefficient of TIC.

From (6), it results that $U_{TIC}(y_r,y_s)\in[0,1]$. It describes a kind of relative error which is very convenient to understand and use. But, when applied to the validation of simulation model directly, the following problems will be appeared.

As shown in Fig. 2(a), $y_r(t)\equiv c_r$, $y_s(t)\equiv c_s$, $t=1,2,\cdots,p$. Meanwhile, $C_r\le 0$ and $C_s>0$ are constant. From (6) results that, $U_{TIC}(y_r,y_s)\equiv 1$. It indicates that the "proximity" of position of y_r and y_s is the worst, which is obviously unreasonable. As shown in Fig. 2(b), $y_r(t)=f(t)$,

$y_{s1}(t) = f(t) + c$, $y_{s2}(t) = f(t) - c$, $t = 1, 2, \cdots, p$, where $c > 0$ is constant. It is easily obtained that $U_{TIC}(y_r, y_{s1}) = U_{TIC}(y_r, y_{s2})$ by the intuitive judgment. However, from (6) results that $U_{TIC}(y_r, y_{s1}) > U_{TIC}(y_r, y_{s2})$.

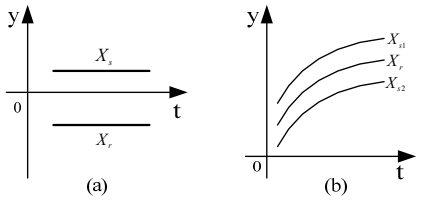


Figure 2. Outputs of reference and simulation in special conditions

The reason why the results above are not consistency is that the relative error of y_r and y_s is calculated by the benchmark of

$$\sqrt{\sum_{i=1}^{p} y_r(i)^2} + \sqrt{\sum_{i=1}^{p} y_s(i)^2} , \tag{7}$$

not by the benchmark of reference outputs.

The measurement model of the "proximity" of position of y_r and y_s is given based on the coefficient of TIC in this paper. The measurement model is as follows.

$$U_{TIC}(y_r, y_s) = \begin{cases} \dfrac{\sqrt{\sum_{i=1}^{p} (y_r(i) - y_s(i))^2}}{\sqrt{\sum_{i=1}^{p} y_r(i)^2}}, & \sqrt{\sum_{i=1}^{p} y_r(i)^2} \neq 0 \\ \sqrt{\sum_{i=1}^{p} (y_r(i) - y_s(i))^2}, & \sqrt{\sum_{i=1}^{p} y_r(i)^2} = 0 \end{cases} \tag{8}$$

$$d(y_r, y_s) = e^{-\lambda_d T(y_r, y_s)} \tag{9}$$

where $\lambda_d > 0$ is the parameter of measurement model for the "proximity" of position. λ_d is given by the area of specific application.

The simulation model is validated by the "similarity" of shape between y_r and y_s, which is depicted by the method of grey relevance coefficient in [8]. The formula is as follows:

$$r(t) = \frac{\min\limits_{t} \Delta(t) + \lambda_s \max\limits_{t} \Delta(t)}{\Delta(k) + \lambda_s \max\limits_{t} \Delta(t)} , \ t = 1, 2, \cdots, p \tag{10}$$

$$s(y_r, y_s) = \frac{1}{p} \sum_{t=1}^{p} r(t) \tag{11}$$

where $\Delta(t) = |y_r(t) - y_s(t)|$, $\lambda_s \in [0,1]$ is resolution coefficient. The range of λ_s is $[0, 0.5]$.

Considering the "proximity" of position and the "similarity" of shape for y_r and y_s based on (9) and (11), the data consistency measurement model of y_r and y_s is given as follows:

$$D(y_r, y_s) = \sqrt{d(y_r, y_s) \times s(y_r, y_s)} \tag{12}$$

B. Feature consistency measurement model

Some features extracted from the reference output and the simulation output, are denoted by c_r and c_s respectively. The relative error is used to describe the difference between them, as shown in (13) [14]. The degree of consistency is obtained by mapping the relative error to interval (0,1] through negative exponential function, as shown in (14) [15].

$$\eta_c = \begin{cases} \dfrac{|c_s - c_r|}{|c_r|}, & c_r \neq 0 \\ |c_s - c_r|, & c_r = 0 \end{cases} \tag{13}$$

$$V(c_s, c_r) = e^{-\lambda_c \eta_c} \tag{14}$$

where $\lambda_c > 0$ is the parameter of consistency measurement model, given by the area of specific application.

IV. INTEGRATED MODEL OF CONSISTENCY

$D(y_{si}, y_{ri})$, $i = 1, 2, \cdots, m$ and $V(c_{sj}, c_{rj})$, $j = 1, 2, \cdots, l$ can be got by using the consistency measurement model above. Furthermore, the consistency of simulation output can be obtained by integrating $D(y_{si}, y_{ri})$ and $V(c_{sj}, c_{rj})$ as:

$$C(Y_s, Y_r) = \sum_{i=1}^{m} \omega_{di} \times D(y_{si}, y_{ri}) + \sum_{j=1}^{l} \omega_{vj} \times V(c_{sj}, c_{rj}) \tag{15}$$

where ω_{di} and ω_{vj} are the weights of data consistency and feature consistency, respectively.

From all above, in order to get integrated model of consistency, the important thing is to determine the weights of $D(y_{si}, y_{ri})$, $i = 1, 2, \cdots, m$ and $V(c_{sj}, c_{rj})$, $j = 1, 2, \cdots, l$. However, the correlation exist between $D(y_{s1}, y_{r1}), \cdots, D(y_{sm}, y_{rm})$, $V(c_{s1}, c_{r1}), \cdots, V(c_{sl}, c_{rl})$. Thus, the weight of each index is determined based on the correlation analysis in [16].

Due to the reference output samples are usually less, it assumes that there is only a single sample here. Each component of the output time series is expressed as $y_{ri} =< y_{ri}(1), y_{ri}(2), \cdots y_{ri}(p) >$, $i = 1, 2, \cdots, m$. Even when a few samples of the reference output exist, they should be averaged into a single time series. Simulation output samples are usually easy to obtain. It is assumed that q samples of each output component are $y_{si}^j =< y_{si}^j(1), y_{si}^j(2), \cdots y_{si}^j(p) >$, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, q$.

Furthermore, z_i , $i = 1, 2, \cdots, m + l$ denotes the data consistency and feature consistency between simulation output and the reference output. z_i^j , $i = 1, 2, \cdots, m + l$, $j = 1, 2, \cdots, q$ denotes the data consistency and feature consistency between the j th sample of simulation output and the reference output. To determine the weight of z_i ,

$$\bar{z}_i = \frac{1}{q} \sum_{j=1}^{q} z_i^j , \ i = 1, 2, \cdots, m + l \tag{16}$$

$$s_{ii} = \frac{1}{q}\sum_{j=1}^{q}(z_i^j - \overline{z}_i)^2 \ , \ i = 1,2,\cdots,m+l \qquad (17)$$

$$s_{ij} = \frac{1}{q}\sum_{h=1}^{q}(z_i^h - \overline{z}_i)(z_j^h - \overline{z}_j) \ , \ i = 1,2,\cdots,m+l \qquad (18)$$

Furthermore,

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \ , \ i,j = 1,2,\cdots,m+l \qquad (19)$$

z_i and other indexes of the multiple correlation coefficient can be obtained as follows:

$$R_i = \frac{1}{m+l-1}(\sum_{j=1}^{m+l} r_{ij}^2 - 1) \ , \ i = 1,2,\cdots,m+l \qquad (20)$$

The degree z_i "included" by other indexes can be reflected by the multiple correlation coefficient. If the larger R_i became the worse the independence of z_i, whereas the better. So the weight of z_i can be determined as follows:

$$a_i = \frac{\max\limits_j R_j}{R_i} \ , \ i,j = 1,2,\cdots,m+l \qquad (21)$$

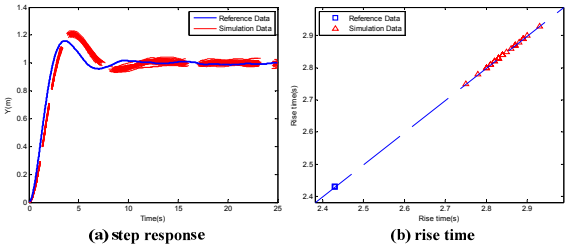
$$\omega_i = \frac{a_i}{\sum_{j=1}^{5} a_i} \ , \ i,j = 1,2,\cdots,m+l \qquad (22)$$

Thus the integration model of consistency can be got as:

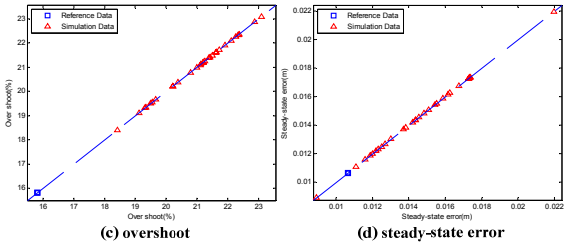
$$C(Y_s, Y_r) = \sum_{i=1}^{m+l} \omega_i \times \overline{z}_i \qquad (23)$$

V. APPLICATION

The object of application studied in this paper is a model of servo-control system. The reference output is the step response of the actual system; the simulation output is 30 times the running output of the simulation model, as shown in Fig. 3(a). There are two ways to validate the credibility of simulation model using the classical methods: 1) Analyze the consistency of step response between the actual system and simulation model directly; 2) Extract rise time, overshoot and steady-state error from step response of the actual system and simulation model respectively, then analyze the features consistency between them. However, no matter which way is used above, the consistency between the simulation output and the reference output is portrayed from one-sided. In addition, the correlation between multiple data consistency and features consistency was not considered by the classic methods.



(a) step response (b) rise time



(c) overshoot (d) steady-state error

Figure 3. Outputs of reference and simulation in special conditions

TABLE I. THE ANALYSIS RESULTS OF DATA CONSISTENCY AND FEATURES CONSISTENCY

N	C_y	C_t	C_p	C_e
01	0.852	0.913	0.894	0.822
02	0.838	0.919	0.824	0.730
03	0.829	0.910	0.865	0.733
04	0.852	0.910	0.869	0.769
05	0.851	0.902	0.885	0.840
06	0.841	0.923	0.854	0.799
07	0.860	0.931	0.846	0.922
08	0.845	0.910	0.900	0.909
09	0.859	0.921	0.829	0.926
10	0.855	0.915	0.832	0.772
11	0.843	0.936	0.815	0.783
12	0.838	0.927	0.838	0.846
13	0.831	0.927	0.819	0.939
14	0.837	0.919	0.845	0.730
15	0.867	0.910	0.890	0.833
16	0.848	0.923	0.837	0.895
17	0.834	0.919	0.800	0.732
18	0.858	0.920	0.835	0.796
19	0.845	0.913	0.895	0.840
20	0.855	0.919	0.814	0.931
21	0.837	0.921	0.870	0.751
22	0.840	0.921	0.832	0.980
23	0.866	0.925	0.849	0.917
24	0.839	0.912	0.812	0.733
25	0.857	0.908	0.921	0.862
26	0.840	0.913	0.843	0.957
27	0.852	0.925	0.841	0.812
28	0.873	0.927	0.794	0.866
29	0.848	0.917	0.843	0.945
30	0.821	0.921	0.888	0.588

For the problems above, this paper not only considers the data consistency of step response, but also considers the consistency of rise time, overshoot and steady-state error. By analyzing the correlation between them, the respective weights are determined. The consistency analysis results based on (7) to (13) are shown in Table 1, where λ_d, λ_s, and λ_c are taken as 0.5; N, C_y, C_t, C_p and C_e in these equations represent the number of simulation runs, the consistency degree of step response, the consistency degree of rise time, the consistency degree of overshoot, and the

 ISBN: 978-1-61208-358-2

consistency degree of steady-state error respectively. Furthermore, the weight of each index obtained by (15) to (21) shows as follows:

$$\omega = [\omega_y, \omega_t, \omega_p, \omega_e] = [0.1523, 0.1110, 0.2126, 0.5241] \quad (24)$$

For the 30 samples of C_y, C_t, C_p and C_e, the average value can be obtained as follows:

$$[\overline{C}_y, \overline{C}_t, \overline{C}_p, \overline{C}_e] = [0.7607, 0.7459, 0.7367, 0.8462] \quad (25)$$

Finally, according to (22), the consistency analysis result of simulation output can be obtained as follows:

$$C(Y_s, Y_r) = [\omega_y, \omega_t, \omega_p, \omega_e] \times [\overline{C}_y, \overline{C}_t, \overline{C}_p, \overline{C}_e]^T = 0.7988 \quad (26)$$

The analysis result shows that compared to the consistency of the reference output in the instance, the consistency of the simulation output consistency is better. It indicates that the simulation model is credible. By observing the simulation output and the reference output from Fig. 3(a) to 3(d), it can be concluded that they have good consistency. The results of this paper are identical with this conclusion. Thus, the effectiveness of the method in this paper is verified.

VI. CONCLUSION AND FUTURE WORK

This paper focused on the study of the validation of simulation model. Our goal is to obtain the consistency degree of simulation output and reference output by considering the data consistency and feature consistency simultaneously. In this paper, three conclusions are obtained as follows: 1) The new research thought of simulation model validation was given. 2) The measurement model of data consistency integrating the "proximity" of position and "similarity" of shape and the measurement model of feature consistency based on relative error are presented. 3) The method of how to integrate two types of consistency measurement model was proposed. However, the scarcity of the reference output may exist in multi-output dynamic simulation. So, in future research, it is planned to concentrate on how to integrate with the expert knowledge to validate simulation model.

ACKNOWLEDGMENT

This paper is supported by the Fundamental Research Funds for the Central Universities (HIT.NSRIF.2015035).

REFERENCES

[1] R. G. Sargent, “Verification and validation of simulation models,” Journal of Simulation, vol. 7, 2013, pp. 12–24.

[2] F. Y. Min, M. Yang, and Z. C. Wang, “Knowledge-based method for the validation of complex simulation models,” Simulation Modelling Practice and Theory, vol. 18, 2010, pp. 500–515.

[3] A. G. Biggs and A. R. Cawthorne, “Bloodhound Missile Evaluation,” Journal of the Royal Aeronautical Society, vol. 66, 1962, pp. 571–598.

[4] G. S. Fishman and P. J. Kiviat, “The Analysis of simulation Generated Time Series,” Management Science, vol. 3, 1967, pp. 525–557.

[5] C. F. Hermann, “Validation Problems in Games and Simulations with Special Reference to Models of International Politics,” Behavioral Science, vol. 12, 1967, pp. 216–231.

[6] J. L. Mckenny, “Critique of Verification of Computer Simulation Models,” Management Science, vol. 14, 1967, pp. 55–59.

[7] N. A. Kheir and W. M. Holmes, “On Validating Simulation Models of Missile Systems,” Simulation, vol. 30, 1978, pp. 117–128.

[8] H. L. Wei and Z. W. Li, “Grey Relational Analysis and Its Application to the Validation of Computer Simulation Models for Missile Systems,” Systems Engineering and Electronics, vol. 19, 1997, pp. 55–60.

[9] M. J. Damborg, “An example of error analysis in dynamic model validation,”. Simulation, vol. 44, 1985, pp. 301–305.

[10] S. K. Liu, X. T. Liu, and W. Zhang, “Fixed Quantity Evaluation to Reliability of Simulation System with Similar Degree,” Journal of System Simulation, vol. 14, 2002, pp. 143–145.

[11] R. G. Sargent, “Verification and validation of simulation models,” Proceedings of the 2011 Winter Simulation Conference, 2011, pp. 183–198.

[12] O. Balci, “Verification, validation, and certification of modeling and simulation applications,” Proceedings of the 2003 Winter Simulation Conference, 2003, pp. 150–158.

[13] J. P. C. Kleijnen, “An overview of the design and analysis of simulation experiments for sensitivity analysis,” European Journal of Operational Research, vol. 164, 2005, pp. 287–300.

[14] J. M. Morales and J. Perez-Ruiz, “Point estimate schemes to solve the probabilistic power flow,” IEEE Transactions on Power System, vol. 22, 2007, pp. 1594–1601.

[15] T. Mei, C. H. Yang, G. B. Wang, and et al., “A new method for estimating rock joint size,” Chinese Journal of Rock Mechanics and Engineering, vol. 27, 2008, pp. 3503–3508.

[16] Y. C. Song, Y. Y. Zhang, and H. D. Meng, “The research based on euclid distance with weights of clustering method,” Computer Engineering and Applications, vol. 43, 2007, pp. 179–180.

 ISBN: 978-1-61208-358-2

Early Detection of Critical Faults Using Time-Series Analysis on Heterogeneous Information Systems in the Automotive Industry

Thomas Leitner*, Christina Feilmayr†, Wolfram Wöß‡
Institute for Application Oriented Knowledge Processing, Johannes Kepler University Linz
Linz, Austria
Email: *thomas.leitner@jku.at, †christina.feilmayr@jku.at, ‡wolfram.woess@jku.at

***Abstract*—Beside the manufacturing industry's vision of *industry 4.0*, which is about improving the degree of automation and customisability depending on a huge amount of data, the automotive industry increasingly advances the after-sales market collecting more and more information about the car using sensors and diagnostics mechanisms. This information can be used to earlier reveal malfunctions and faults with rising quantity that customers experience in order to reduce the solving time of the problem. Different heterogeneous data sets exist storing data at various approval stages with different data quality. In order to perform the most accurate detection of critical developing faults it is fundamental to use as much data as possible while weighting their impact by assessing their data quality. For detecting critical performing faults as early as possible time series analysis and forecasting methods are used to analyze their course and predict future values. In this research work, a new approach is proposed, which is subdivided in the following four main tasks: (i) evaluation of data quality metrics of different warranty information systems, (ii) analysis and generation of forecasts on univariate time series based on *Auto-Regressive Integrated Moving Average* (ARIMA), (iii) weighted combination of different predictions, and (iv) improvement of the accuracy by integrating prediction errors. This solution can be used in arbitrary fields of application, in which different information sources should be analyzed using data quality metrics and prediction errors to determine critical courses as early as possible.**

Keywords–data mining, time series analysis, data quality metrics, automotive industry

I. INTRODUCTION

Quality - abnormality and cause analysis (Q-AURA – *Qualität - Auffälligkeiten und Ursachenanalyse*) is an application developed in cooperation with *BMW Motoren GmbH*, located in Steyr, Austria, with the purpose to decrease the problem solving time for finding causes of engine faults in automobiles in the after-sales market. The system has different goals for supporting the quality management expert in his daily work; (i) automatically finding significant faults that are developing badly, (ii) providing new useful information about the affected engines, and, (iii) analyzing bills of materials and engine components to find technical modifications that potentially provoke a particular fault. Q-AURA has already been evaluated and is used by the quality management experts every day. The first task of detecting significant faults uses fault information from warranty claims of previous weeks, but takes only those information within a specific approval stage into account that originate from a single data source. The goals to additionally use information of other warranty information systems at other approval stages and to earlier determine significant faults leads to the need of an optimisation of the existing system. Methods should be investigated that help to achieve robust results.

In this paper, an approach is presented that uses time series analysis, forecasting methods, and data quality assessment of different information systems. The paper is organized as follows: Section II discusses the central problem and associated challenges. Section III addresses methods that are necessary for the provided approach, while Section IV gives a detailed description of the proposed technique also explaining the integration into Q-AURA. Finally, Section V covers the conclusion.

II. PROBLEM STATEMENT

The contribution of this research work is a new approach, which consists of four parts, whereas each of them discusses a particular problem. The first one is the development of a *method for validating different information systems, which store partially contradictory, complementary, and redundant information*. The business process that is supported by Q-AURA ranges from the engine development department where new engine generations are developed or existing ones are improved to the after-sales market. In case of a technical fault, the car must be checked at a dealer's workshop, where the problem and information about the fault is sent to the automobile manufacturer. Since BMW sells cars in most countries, and since faults are classified differently in various markets, it is necessary to overcome the discrepancies yielding in a consistent view of all faults that occur. Different information systems exist that contain faults at different acceptance levels. These data sets are evaluated using data quality metrics.

The second task is the development of a *method for detecting critical developing faults as early as possible*. In order to identify critical faults, the trend of the most recent weeks is determined. Since enough values are necessary to provide robust results, there is a time delay between the beginning of the fault and its detection. By using prediction methods this delay can be reduced since future values can be predicted, which can then be used to determine earlier whether a particular course is critical or not. Different prediction methods have been evaluated and the best one was selected.

The third task is the development of a *concept to improve the prediction accuracy using forecasts from different information systems*. As explained above different views on fault and warranty information at different approval stages exist. E.g., while one source contains information that is already accepted by the company, but does not include values that are provoked by the customer, another one contains more faults, but those have not been verified by the company. Therefore, these different sources have to be analyzed separately, which results in separate predictions. By consolidating the forecasts

 ISBN: 978-1-61208-358-2

performed on each individual data set a better result can be achieved. Since the information quality of each individual data source has to be taken into account, the quality scoring is used to influence the weighting to get more accurate results.
The fourth task is the *integration into Q-AURA and verification of the proposed concepts* to demonstrate the improvement in comparison to the current applied approach. It is important to describe the established consolidated Q-AURA system to clearly show the benefit of the new approach.
The resulting approach is a set of methods that enable early detection using a weighted combination of forecasts based on data of multiple, heterogeneous information systems that adjusts its parameters using accuracy metrics of previous iterations and quality metrics of each data source.

III. Related Work

This section covers information about the used methods and gives a detailed description of the concepts the proposed approach is based on. Primarily, two basic concepts are discussed, which are data quality metrics including their assessment and analysis of univariate time series containing forecasting methods.

A. Data Quality

The literature provides a wide range of techniques for data quality assessments as well as definitions and descriptions about data quality dimensions and metrics [1], [2]. A detailed comparison is given by Batini et al. [1]. For a compact summary, the data quality metrics that are used in this approach are described in detail.
Completeness describes if all information in the real world within a particular scope is captured by the information system. In other words, a system is complete if it includes the whole truth. For a database scheme D, we assume a hypothetical database instance d_0 that perfectly represents all the information of the real world that is modelled by D. Further, we assume one or more instances $d_i (i \geq 1)$, where each of them is an approximation of d_0. Now we consider some views, where v_0 is an ideal extension of d_0 and $v_i (i \geq 1)$ are extensions of the instances d_i. Further we define completeness as described by (1). In the considered equation, the absolut values represent the number of tuples [3], [4].

$$\frac{|v_i \cap v_0|}{|v_0|} \tag{1}$$

Soundness (similar to completeness) is also determined by comparing the real world and the modelled instances in the information system. It describes if the information system stores the truth, and nothing but the truth, which means that all modelled information also exists in the real world. Equation (2) shows the definition [3], [4].

$$\frac{|v_i \cap v_0|}{|v_i|} \tag{2}$$

Consistency is a metric that focusses on the structural correctness of the represented data. This means that stored information has to meet some conditions, e.g., existing entries have to be unique (no duplicates) or meet assertions. In the literature, different definitions exist, some of them are similar to soundness in others [1].
Correctness is a metric that measures the semantic validity of data. Stored data is correct if it meets semantic rules. By applying those rules it can be determined if the particular entry is in the correct range or has a valid format, for example. Since functional requirements can change over time, it is important to modify these rules if necessary [5].
Integrity is also defined ambiguous in literature. Some definitions declare integrity as the combination of validity and completeness [6]. In this contribution, integrity is treated as the correctness of connections between data structures like tables or views. This means that the connections between data structures are monitored and if too many, wrong or too few results are calculated than expected this metric is decreased. Sometimes *inter-relation integrity* has a similar definition in literature [1].
The data quality metrics used in this approach were chosen carefully. While completeness and soundness measure if the quantities of the basic structural components are correct, integrity proves if the connections between them are valid. Consistency and correctness determine if the entries are semantically correct and in the specified value range.

B. Analysis of Univariate Time Series

This section focusses on time series analysis especially univariate ones. Time series analysis is also a very popular research field since a wide variety of applications exists, ranging from stock analysis and calculations concerning demography to sunspot observations. Forecasts are connected closely to time series analysis, since it is the prediction of future values of a known time series. A popular example are weather forecasts, where former observations are known and based on them (and physical law of course) future values are predicted [7].
The proposed approach focusses on univariate time series, which are time series that solely depend on one variable [7], [8]. Different methods exist for modelling and forecasting univariate time series, under them *Box-Cox Transformation, ARMA errors, Trend, and Seasonality* (BATS) and *TBATS*, *Simple Exponential Smoothing* (SES), *feed forward neural networks with a single hidden layer* (NNETAR), *Croston's method* (CROSTON), and *Auto-Regressive Integrated Moving Average* (ARIMA) [9], [10], [11].These different methods were compared with each other, and the best one was chosen for the proposed approach, which is ARIMA. The comparison was done using example data. For the evaluation which method facilitates the best results, *Goodness-of-Fit* measures (e.g., *Mean Percentage Error*, *Mean Absolute Percentage Error*) were used for comparing the fitting of the curve of each method, while the *Diebold-Mariano Test* was used for comparing the accuracy of the predictions [12].
Nevertheless, time series analysis results in meaningful outcomes if enough values are present, so that prediction methods can be applied.

IV. Improving Early Detection of Significant Faults in Quality Management

This chapter describes the Q-AURA analysis process, the invented concept and its integration into Q-AURA. Q-AURA is a system that monitors faults gathered from warranty information systems, where the data originates at the different car

 ISBN: 978-1-61208-358-2

dealers. The addressed business process is depicted in Figure 1 and encompasses phases from the early development of an engine to the after-sales market.

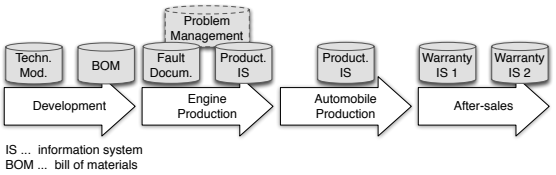


Figure 1: Flow chart of the business process relevant to Q-AURA

It is shown that warranty and fault data of the after-sales market is spread over more than one information system. Since these information systems store partially different data, their integration and combination can provide additional information for determining which faults are developing badly and, thus, have to be investigated further. The figure also presents the involvement of different data sources throughout the whole process in order to get the necessary information of engine components and technical modifications.

A. Q-AURA Approach

First the existing Q-AURA approach is described to explain the underlying analysis method and how the information is processed. Q-AURA provides different steps, each of them is necessary to modify the information in such a way that, finally, potential technical modifications that may cause a particular fault can be determined. Figure 2 illustrates all six steps.

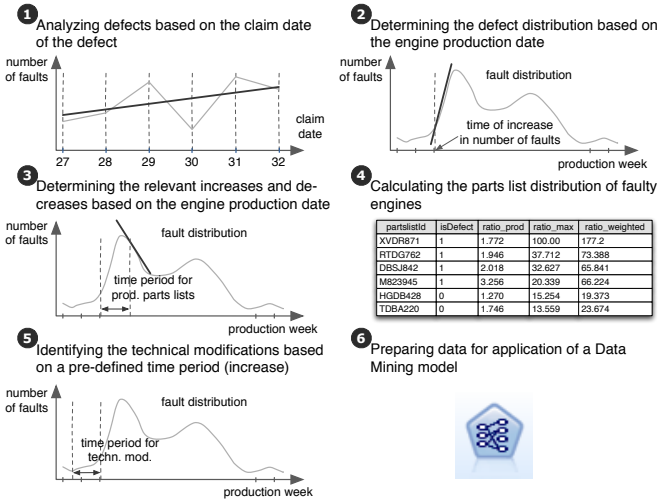


Figure 2: Q-AURA process in detail

The information base for the first step is a set of warranty claims of the last two years from cars produced in the last three years (cf. Figure 2-1). The boundaries were set carefully to determine those cars with corresponding engines that influence the ongoing development process. In the currently used system, faults of warranty claims of the latest six weeks are used to identify current problems with high significance. In order to determine whether the fault is significant or not, a regression analysis is applied [13]. Three different approaches to regression analysis were tested and evaluated containing convex functions, smoothing functions, and a straight line. The evaluation reveals that the straight line approach for regression enables the best results. The *gradient*, *mean value*, and *coefficient of determination* of the regression line are calculated to measure the characteristics of the applied regression. Thresholds, which have been defined and evaluated with the quality management experts, are used to determine if a fault is significant. Significant classified faults are analyzed further. In the second step (cf. Figure 2-2), the production week histogram of engines having a particular fault within the last two years is calculated containing cars produced over the last three years. Next, a normalisation of the fault amount by the total number of produced engines belonging to the same class (cars with the same car brand, fuel type, and engine type) is performed. Afterwards, a 5-point smoothing function is applied to remove the outliers. Significant increases of the resulting smoothed curve are identified to determine those engine production dates, where the ratio between faulty engines and the amount of produced engines of the same type is rising meaning that something changed (e.g., due to a technical modification). In the next step (cf. Figure 2-3), significant decreases are determined. Afterwards, the bill of materials (BOM) distribution of the faulty engines of each period bounded by a significant increase and its subsequent decrease is calculated. The BOM distribution is normalized by the production volumes in order to determine those BOMs that have a bad ratio and, therefore, most likely contain a causing technical modification (cf. Figure 2-4). In the next step (cf. Figure 2-5), the technical modifications of critical BOMs are limited by those that were set operational in a time period before and after the significant increase. Finally, the critical technical modifications are determined using two different methods (cf. Figure 2-6). The first one selects modifications, which most of the critical BOMs contain. The second method uses association rules, in this case the *Apriori* algorithm, for the same task [14], [15].

Q-AURA was already set operational and is used by the quality management experts for their daily work. An evaluation has been done stating that a significant benefit was achieved.

B. Concept of Improving Early Detection

It is obvious that faults that occur anytime during the development and production process should be detected as early as possible. It is not only important because of financial matters, there is also a disadvantage for the reputation of the brand and, consequently, for the company as well. Because of the fact that the analysis of causes can be a time consuming task, even an acceleration by a single day is a big advantage. The proposed approach focusses on detecting faults that develop critically earlier, which means, that Q-AURA can detect potential technical modifications earlier. This results in a reduction of the problem solving time. The proposed improvement uses four main concepts, which are, (i) assessing data quality of information systems storing warranty information, (ii) analyzing and forecasting univariate fault time series, (iii) combining and weighting different predictions and, finally, (iv) evaluating the prediction accuracy followed by an adjustment of the weighting of information sources.

The overall concept is depicted in Figure 3 and consists of the components validator, predictor, combiner, and controller. Each of them has a specific task in the process. First (cf. Figure

 ISBN: 978-1-61208-358-2

3-1), the validator's task is to estimate the data quality of the various information sources. This is done by using different quality metrics. The quality metrics that are used by the presented approach are completeness, soundness, consistency, correctness, and integrity. An overall metric is calculated by a combination of the different factors. Afterwards, the predictor computes forecasts based on the fault and warranty entries of the different information systems (cf. Figure 3-2). As a result an enhanced time series exists containing the predicted value. Linear regression as it is currently used by the Q-AURA approach is applied on the most recent six weeks of the fault time series (containing the newly predicted value). Consequently, the characteristic metrics of the regression analysis performed on each information base is calculated. Subsequently, the combiner's task is to derive the overall prediction whether the fault is significant or not using the parameters of the predictor (cf. Figure 3-3). In order to verify the significance of the prediction, the quality metric of each information system is used. Finally, the controller is necessary to determine the accuracy of the different forecasts (cf. Figure 3-4). This is done by comparing new information system entries of the next week with the predicted values of the predictor. The difference (prediction error) is another weighting factor, which is integrated into the combiner's method. In the next iteration, the combiner uses the newly computed weighting of the controller to adjust the impacts of the information systems. Next, the different concepts of the proposed components are explained in more detail.

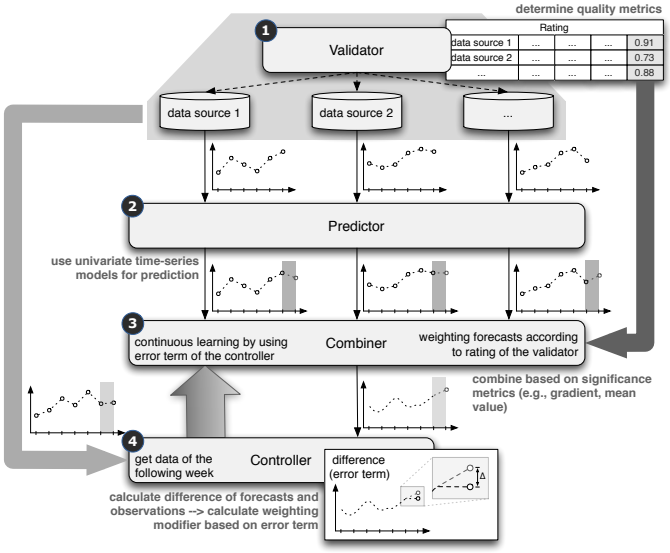


Figure 3: Concept for improving early detection

1) Validator: The task of the validator component is determining the overall data quality of the different information systems. Different understanding exists about data quality metrics, therefore, those used are explained. The presented approach uses the quality metrics completeness, soundness, consistency, correctness, and integrity.

Completeness determines if all the information of the real world is captured in the particular data source. The presented approach focusses on closed world assumption which means that any information that is not modelled, is treated as not existing. This means that *Null* values are treated as missing values. Since it is nearly impossible to determine all instances of the real world, an assumption in the proposed approach is made. The real world is approximated by consolidating all instances of the different information systems. The method, which is used for determining the completeness quality metric is relation completeness [2].

Soundness addresses the difference between the real world and the entries of the data sources. This metric indicates if the particular information system stores false values. For an approximation of the real world a combination of the data stored in the different information systems is used. As described in Section III-A, this value is defined as the ratio of the intersection of the information source and the real world and the number of entries in the information source.

Consistency is also a crucial data quality metric and measures the goodness of the entries in the information systems, which is done by proving if duplicates exist or if entries are defined ambiguously. If constraints, referential integrity, and primary keys are applied correctly, this metric can be increased in many cases.

Since consistency does not prove if inserted values are valid, the quality metric correctness is used. This one is very difficult to check, since a software can not automatically prove on its own if a value is correct. Such functional requirements have to be integrated explicitly by implementing concrete rules.

The last quality metric that is applied by the proposed approach is integrity. This metric is assessed by inspecting the join operations between data structures of an information source. This means that if a master data table contains too few entries for a corresponding transaction data table, then this value is decreased.

The overall quality metric is calculated as product of the different single quality metrics, because each feature also influences the other quality metrics (see Figure 4).

data source	completeness	soundness	consistency	correctness	integrity	quality
data source 1	0.85	0.99	0.91	1.00	0.98	0.75
data source 2	0.94	0.94	0.97	0.98	0.96	0.80
data source 3	0.98	0.96	0.89	0.92	0.85	0.65
...	...	...	...	...	...	...

Figure 4: Exemplary results of the validator component

2) Predictor: Various time series methods exist to model the behaviour of univariate time series as explained in Section III-B. These methods were applied on warranty and fault information, which is used in Q-AURA. *Goodness-of-Fit* measures and the *Diebold-Mariano Test* were taken for the comparison. The ARIMA method was chosen, since it delivered the best results.

Linear regression is applied on the most recent six weeks of the time series including the predicted value of the next week to determine its characteristics. A straight line is used for fitting, and the features *gradient*, *mean value*, and *coefficient of determination* are determined. The first two parameters can be calculated using following equation (see Equation (3)):

$$y = k * x + d \tag{3}$$

The parameter x determines a point in time on the x-axis of a histogram, while y is the corresponding measured value.

 ISBN: 978-1-61208-358-2

Together these two values determine the coordinates of a data point on the line of regression. The value k is called the gradient and describes the increase between two data points. d is called the offset and is equal to the y-value at the point $x = 0$. The mean value $\bar{y}$ is the average value of the measured points over the six weeks period. In order to describe the steadiness of the curve the *coefficient of determination* is determined [13]. If the regression line depends only on one variable as it is the case if a straight line is used, the coefficient of determination is equal to the square of *Pearson's Correlation Coefficient* r_{xy}^2 (see Equation (4)) [16]:

$$R^2 = r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \tag{4}$$

After the predictions were performed on each data source, it results in an output as depicted in Figure 5.

data source	gradient	mean value	coeff. o. det.
data source 1	1.92	12.34	0.32
data source 2	2.45	32.91	0.19
data source 3	0.64	24.54	0.98
...	...	...	...

Figure 5: Exemplary results of the predictor component

3) Combiner: In the next step, the combiner uses the previously calculated metrics and determines an overall significance value, which specifies whether the significant fault is critical or not. As explained previously, the characteristic metrics of a regression line can be assessed by using threshold values that have been investigated with experts of the quality management department. There are two different methods available for implementing the combiner component, both based on weighted voting. Weighted voting uses weighting factors to determine the impact of each single data source.

- *Combination of the characteristic metrics*: This option uses the gradient, mean value, and coefficient of determination for the combination task. The relative differences between the gradients of the various data sources and the defined threshold are calculated. Afterwards, the results are combined using weighted voting. The same procedure is applied to the coefficients of determination of the different data sources. Because of the fact that the mean value ensures that sufficient values exist for a robust Q-AURA determination of technical modifications, it is assessed if the mean value of each source exceeds the given threshold. Afterwards, weighted voting is applied on these assessment results.
- *Combination of the significance results*: Before the combination is performed, the characteristic metrics of each data source's regression line is assessed whether the particular fault is significant or not. This results in a significance value for each data source, which is integrated using weighted voting.

Since the first variant (combination of the characteristic metrics) determines the outcome on a more fine-grained basis, it is used in the presented approach.
In the proposed approach, the weighting used by the combiner consists of two components. The first one was already explained earlier and is calculated by the validator component, representing the overall data quality of the various information systems. The second component is explained next and is the accuracy determined by the controller component using the prediction error.

4) Controller: The task of the controller is assessing the prediction quality of the data sources. Since the prediction in the proposed approach is always a one step forecast, which means that only one future value (one week ahead) is calculated, the validation can be done in the following week. Thus, the grading and impact of the actual week's computation is based on the controller's results of previous weeks. In order to reduce the impact of outliers, the assessment of the prediction accuracy is not only based on the last single week. The accuracies of the previous weeks will also influence the calculation, which is achieved by integrating the accuracy metric of the previous value. Since the accuracy value of the previous week was calculated using the observation of the last week and the accuracy of the week before, all the previous accuracies influence the actual value, but with decreasing impact (Equation 5).

$$p_t = \frac{1}{2} * (\frac{|f_{t-1} - x_t|}{f_{t-1} + x_t} + p_{t-1}) \tag{5}$$

The p values in the equation represent the calculated prediction accuracies, while f_{t-1} represents the forecasted value ($t-1$ shows that it is the forecasted value of the previous week), which is assessed using the actual week's observation x_t. The first term considering the accuracy of the actual week (forecast of the previous week) is a modification to the *Symmetric Mean Absolute Percentage Error (sMAPE)*. Since the highest possible percentage of the standard sMAPE is 200%, a factor is applied to reduce this to 100% for better applicability. After the prediction accuracy values were calculated for each data source, the results can be used for the combiner's computation [17], [18], [19].

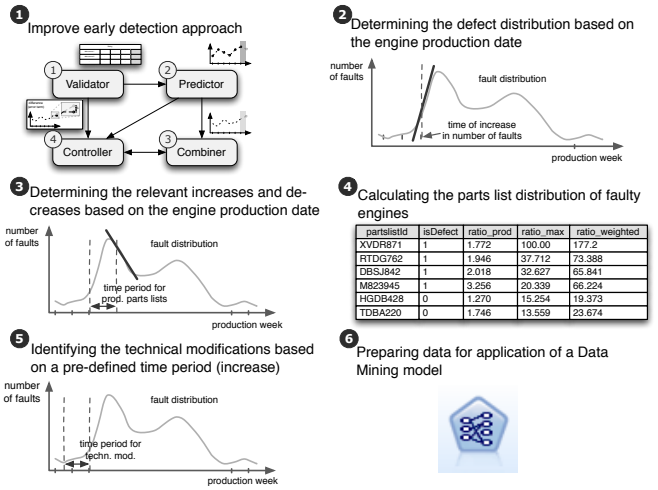
C. Q-AURA Integration

This section describes the integration of the improved early detection concept into the currently used Q-AURA approach, which was explained in Section IV-A. Figure 6 shows the overall concept with improved early detection of significant faults. Instead of applying the six week regression analysis the whole process described in Section IV-B is performed to determine whether a fault is significant or not (cf. Figure 6-1). Afterwards, the final outcome is each fault's significance factor. The rest of the Q-AURA approach remains the same (cf. steps 2-6 Figure 6).

Currently, the approach is applied on the data sources of BMW and is still in testing phase. The actual results are very promising, but more extensive investigation have to be done before detailed results can be published.

V. CONCLUSION

The proposed approach in this research work builds upon an existing system Q-AURA, that is already successfully used by the quality management department in their day-to-day work. Q-AURA is a system that monitors engine faults and



partslistId	isDefect	ratio_prod	ratio_max	ratio_weighted
XVDR871	1	1.772	100.00	177.2
RTDG762	1	1.946	37.712	73.388
DBSJ842	1	2.018	32.627	65.841
M823945	1	3.256	20.339	66.224
HGDB428	0	1.270	15.254	19.373
TDBA220	0	1.746	13.559	23.674

Figure 6: Q-AURA integration of the proposed approach

determines actual problems that develop badly. Afterwards, an automated computation of these faults is performed to find interesting patterns about the cars, resulting in potential technical modifications that may be the cause of faults. Currently, Q-AURA uses linear regression on time series of faults that occurred in the previous six weeks of automobiles that were produced in the last three years.

The contribution addressed in this research work is an approach to detect significant (badly developing) faults earlier by combining predictions of univariate fault time series based on after-sales information, which is stored in different databases. In order to get more accurate results, these different forecasts are weighted according to previous prediction accuracies and data quality metrics of the data sets. The developed technique consists of different components, each of which meets a particular challenge. The validator computes the quality metric for each data source by calculating and combining the metrics completeness, soundness, consistency, correctness, and integrity. Afterwards, the predictor analyzes fault time series based on warranty and claim information of each data source resulting in a forecast of the next week. The controller compares the predictions of the previous week with the observations of the actual week and calculates a weighting factor including the accuracy of previous forecasts. Finally, the combiner integrates the different predictions and determines by weighting and consolidating the values whether the particular fault is significant or not. The approach is already applied and the results are very promising.

REFERENCES

[1] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," ACM Computing Surveys, vol. 41, no. 3, July 2009, pp. 1–52.

[2] C. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3] A. Motro and I. Rakov, "Estimating the Quality of Data in Relational Databases," in In Proceedings of the 1996 Conference on Information Quality. MIT, 1996, pp. 94–106.

[4] Motro, Amihai and Rakov, Igor, "Estimating the Quality of Databases," in FQAS, ser. Lecture Notes in Computer Science, T. Andreasen, H. Christiansen, and H. L. Larsen, Eds., vol. 1495. Springer, 1998, pp. 298–307.

[5] Y. Wand and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," Communications of the ACM, vol. 39, no. 11, November 1996, pp. 86–95.

[6] A. Motro, "Integrity = Validity + Completeness," ACM Transactions on Database Systems, vol. 14, no. 4, December 1989, pp. 480–502.

[7] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications: With R Examples, 3rd ed. Springer Texts in Statistics, 2011.

[8] P. S. P. Cowpertwait and A. V. Metcalfe, Introductory Time Series with R, 1st ed. Springer Publishing Company, Incorporated, 2009.

[9] A. M. De Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing," JASA. Journal of the American Statistical Association, vol. 106, no. 496, 2011, pp. 1513–1527.

[10] L. Shenstone and R. J. Hyndman, "Stochastic models underlying Croston's method for intermittent demand forecasting," Journal of Forecasting, 2005.

[11] J. D. Croston, "Forecasting and stock control for intermittent demands," Operational Research Quarterly, vol. 23, no. 3, 1972, pp. 289–303.

[12] F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," Journal of Business & Economic Statistics, vol. 13, no. 3, July 1995, pp. 253–263.

[13] G. U. Yule, "On the Theory of Correlation," Journal of the Royal Statistical Society, vol. 60, no. 4, December 1897, pp. 812–854.

[14] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in Advances in Knowledge Discovery and Data Mining, U. Fayyad and et al., Eds. MIT Press, 1996, pp. 307–328.

[15] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of data, ser. SIGMOD '93. New York, NY, USA: ACM, 1993, pp. 207–216.

[16] M. Mittlböck and M. Schemper, "Explained Variation for Logistic Regression," Statistics in medicine, vol. 15, no. 19, October 1996, pp. 1987–1997.

[17] S. Makridakis and M. Hibon, "The M3-Competition: results, conclusions and implications," International Journal of Forecasting, vol. 16, no. 4, 2000, pp. 451–476.

[18] E. Mangalova and E. Agafonov, "Time Series Forecasting using Ensemble of AR models with Time-Varying Structure," in IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), May 2012, pp. 198–203.

[19] S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, "The accuracy of extrapolation (time series) methods: Results of a forecasting competition," International Journal of Forecasting, vol. 1, no. 2, 1982, pp. 111–153.

 ISBN: 978-1-61208-358-2

GeoTagView: Visualizing Geographic Tags Easily

A Weka customization for geo-visualization aiming to support spatial analysis

Gianpaolo Pigliasco, Gaetano Zazzaro
Soft Computing Laboratory
CIRA (Italian Aerospace Research Centre)
Capua (CE), Italy
{g.pigliasco, g.zazzaro}@cira.it

***Abstract*— This paper presents a Weka extension called GeoTagView able to quickly and easily display the results of data analysis on a geographical map. After installing GeoTagView, a shape file can be loaded and the results of the analysis are displayed in a separate window. The shape file can be achieved by a software for spatial ETL (Extract, Transform & Load) as GeoKettle. The paper also presents a case study concerning the representation of the levels of pollution from the landfills of waste on the map of Campania Region (in the Southern of Italy). The levels were obtained by a clustering algorithm (k-means).**

Keywords*— *Map Visualization; Spatial Analysis; Clustering; Thematic Map; Big Geographical Data; Weka.

I. INTRODUCTION

During the last decades, large amount of geo-spatial data have been, and continue to be, collected in various applications like geographical information system (GIS), computer cartography, environmental planning, using modern data acquisition techniques such as GPS, high resolution remote sensing, etc.

The scope, coverage, and volume of digital geographic datasets are growing rapidly. Complex scientific and social questions could get responses by means of open availability of huge amount of data with a higher spatial, temporal, and thematic resolution, which could be referred to as *Big Geographical Data*.

Simultaneously, over the last years there has been much progress in knowledge discovery, including the development of new techniques for exploring large, heterogeneous geographic datasets.

Geographic representation is the integration of cartography and scientific visualization aimed to explore geographic data and communicate geographic information to private or public audiences. Major geographic visualization tasks include feature identification, feature comparison, and feature interpretation.

Geovisualization concerns the development of theory, methods, and tools for the visual analysis and presentation of geographic data (i.e., any data with geographic information). Clustering visualization consists of aggregating data items to a relative small number of clusters, visualizing the clusters instead of data items, and then providing details (data items) for each cluster upon user request. Maps are essential for visualizing geographic patterns. For example, two different clustering methods often produce different clusters from the same data due to different searching strategies or underlying constraints. It would be useful and often critical to be able to compare the results of such competitive methods, find commonalities, examine differences, crosscheck each other's validity, and thus better understand the data and patterns.

Weka is able to offer support in the entire experimental process of Knowledge Discovery, from the preparation of the input data, to statistical evaluation of learning schemes produced, including the visualization of the input data and the result of processing. Its main strengths lie in the area of classification, therefore all the latest machine learning approaches, along with the more established, have been implemented in a basic, object-oriented structure, and developed in Java. Moreover, there have also been implemented regression algorithms, association rules and clustering.

Many open-source software projects use or, in some way, take advantage of Weka workbench for their aims [11]. However,none of these projects can display geospatial data by importing a shapefile. In this work, we exploit it in order to present a geographic perspective supporting and easing the cluster analysis of threats to health due to a widespread wrongful practice into the surrounding area of Naples and Caserta [3][4][8]. In order to determine the critical towns from urban pollution point of view due to waste disposal sites, we applied a clustering algorithm to assign to each town a hazard index. Furthermore, in order to assign a scale of dangerousness, the index determined was compared with that calculated by the formula domain.

In order to obtain the levels of pollution, the features analyzed by the algorithm of clustering described the types and the dangerousness of landfills in the municipalities of interest, the number of landfills, the percentage of people in impact areas and the environmental exposure index.

In the rest of the paper, we reveal how we conducted our analysis. In particular, in Section II, we briefly illustrate the objectives we set for this work. In Section III, we introduce the software toolkit used for the analysis and the possibility to be extended by providing your own code in a customized release. Afterward, in Section IV, we describe the programmatic specifics for extending Weka and, in Section V, we sketch out the library Geotools to read,

 ISBN: 978-1-61208-358-2

manipulate, analyze and display geographic dataset. Finally, in the section VI we propose a case study to which we successfully applied the customized toolkit.

II. GOALS

According to the specific objective to reuse and integrate analytic capabilities available with open-source software tools, our effort has been directed towards the building of an analysis system for geographic data based on the suite of free tools for Data Mining named Waikato Environment for Knowledge Analysis (from the homonymous university in New Zealand), currently known in the academic world with the acronym Weka.

Weka is a tool for knowledge analysis, through which an expert in a particular field can use machine learning techniques to automatically extract useful information from large data sets. In this paper, we bring together state of the art in the field of machine learning algorithms and tools for data processing.

III. ASPECTS OF SOFTWARE INTEGRATION WITH THE TOOLKIT WEKA

There are a number of software projects that make use of Weka or its algorithms, allow data in ARFF format to be processed, or enable access to Weka functionality from other programming environments (e.g., Mathematica, R, and Matlab interfaces, as well as Python, or Ruby libraries), or stress a specific algorithm for a peculiar branch of knowledge (e.g., Kea for automatic keyphrase extraction [12]). A list of projects related to Weka can be found in the WekaWiki [11].

In this case, we are interested in improving Weka by extending it with functionalities such as treatment of geographic data, since it does not natively support this type of analysis.

In order to do the integration, we planned on developing our own code, using agile design methodologies.

Our research work was inspired by a publication [1] which proposed a framework for interoperable Spatial Data Mining, and even realized a module [2] that fully integrates itself with Weka to facilitate the preparation of complete spatial datasets (Geographic Data Preprocessing).

IV. PROGRAMMING GEOTAGVIEW PLUG-IN

Starting with the version 3.4.4, one can extend the capacity of Weka to use the class dynamic discovery at run-time. In some versions (3.5.8, 3.6.0), this feature was not enabled by default as it is a bit slow in the initial loading and it does not work in environments that do not require setting a CLASSPATH variable (for example, for the application servers). However, later versions (3.6.1, 3.7.0) were enabled again for dynamic discovery, since Weka can distinguish between being a standalone application or just be run in an environment without CLASSPATH.

From the version following the 3.5.5, the same main user interface (Graphic User Interface) of Weka provides a mechanism to extend it adding items to the main menu (see Figure 1) without having to change the code of the related class. Taking advantage of the automatic discovery of classes, it will show all the entries corresponding to components in the package specified by a properties file.

There are only two requirements to be met so that the components can be included in the main menu (Extensions item):

- Developers have to implement the interface Weka.gui.MainMenuExtension;
- The packages they reside in must be listed in the file GenericPropertiesCreator.props, under the entry named as the interface above mentioned.

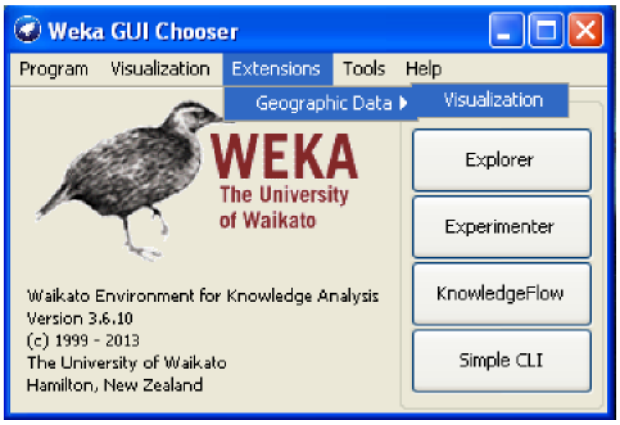


Figure 1. Menu of the main user interface of Weka before and after installed GeoTagsView

The structure of the file is made up of key-value pairs (entry) separated by the equal sign. The value is a sequence of packages separated by a comma.

Figure 1 shows how to access the GeoTagView plug-in by *Extensions* tab menu which can be found after its installation.

V. A THEMATIC MAP FOR WEKA

Geo-referenced databases provide wide opportunities for integration: with GIS we can arrange several geographic datasets of a region in a single database using record linkage based on the location of attributes found. This greatly enriches the statistical analysis because the resulting dataset may contain potential information that none of the starting datasets individually holds. As an example, we can cite the case study of public health, where data related to a particular disease (e.g., from a cancer or birth defects registry in a given region) can be mixed to obtain demographic information on people get involved with the study or relevant information about environmental risks (such as, for example, punctual detection of pollutants, land exploitation, presence of harmful substances related to the substance decomposition in water treatment, and so on).

Moreover, one can conduct a more accurate investigation related to spatial relationships, or other patterns not explicitly observable in spatial databases. The collection of data in relation to the territory allows the production of maps with suitable thematic content representing spatial dynamics of

 ISBN: 978-1-61208-358-2

interesting events. These thematic maps can be used to find out whether the spatial distribution of a phenomenon is concentrated, dispersed, or random. By doing this, one can identify at a glance any territorial concentration statistically significant (spatial cluster) described by similar values concerning the phenomenon addressed and, on the strength of that, generate inferences so as to highlight spatial correlations among observed data.

As we said, we chose to make use of the toolkit Weka, which already includes a number of algorithms for pre-processing, classification, clustering, association rules extraction and data visualization.In this case, however, our aim was to extend viewing capabilities with a new feature, no longer limited to purely nominal data (stated in the native format ARFF – Attribute- Relation File Format), but able to manipulate datasets containing geometric elements within. These will be the subject of our visualization. In effect, geometries correspond to geographic objects whose instance attributes can be variously analyzed in Weka, maybe through a classification or a division into several clusters, so the final representation will be a thematic map showing the result of all analysis activities conducted remaining in the same environment.

There are various techniques to generate a thematic map. The most common way is the so-called chorochromatic, or choropleth, map, which can describe the variability of data under observation through different colors, showing how a phenomenon measure varies within a geographical area in terms of density, percentage, average value, etc.

For storage and exchange of geographic data the *de facto* standard (also used by some of the most important institutions that deal with spatial data) is the Shapefile format, defined in the early 1990s by the Environmental Systems Research Institute, Inc. (ESRI), which ensures the possibility to deal with simple vector data with attributes and, therefore, the ability to record location, shape and information associated with geometric/space entities. This format has become particularly important because it meets the OpenGis Consortium to which Esri acceded.

A shapefile is considered to be a single file, but it is actually a set of several files (of which three are mandatory to store the core data), that simply store the primitive geometric data types of points, lines, and polygons. By themselves, these primitives, called "features", are not sufficient because they are missing of any properties that specify what these primitives represent. Hence, a table of records will store /attributes for each primitive shape in the shapefile. Shapes together with data attributes can create infinitely many representations about geographic data, from which in turn comes the power and accuracy of geospatial analysis that can be done.

In order to achieve in an ”agile manner” an extension able to load and represent geographic features, we chose to make reference to a software library for Java that can provide all the necessary support. Among several possibilities, the most common solution is represented by GeoTools, licensed under the GNU Library General Public License version 2.0 (LGPLv2) [5]. In detail, having this a weight (in terms of size for the storage) not quite negligible, to prevent the plug-in becomes predominant compared with the software to extend, we thought to use a version a bit “dated” of the same library, with a reduced number of functions, but sufficient for our need of just visualizing a simple map.

Figure 2 shows the selection of data fields to make features stand out in the map (eventually, how many classes for the shade) and to give a tip when the mouse passes over.

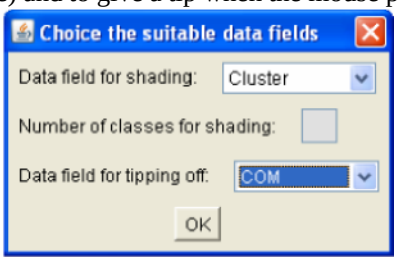


Figure 2. The window for the criteria to paint the map

In addition, it was needed to recompile the entire library in a new package and write a class implementing the Weka interface MainMenuExtension for coordinating the extension behavior.

With these tools we can access data stored in the ESRI shapefile format and use the color associated with geographical elements to represent the increase or decrease of numerical data aggregated by geographical area.

For example, once the levels of environmental pollution of a certain geographic area are known, we could get a map reassuming the result of a spatial clustering based on the cancer incidence, in order to discover possible correlations between the two phenomena.

VI. Case Study

The implemented plug-in was used to obtain cluster representations useful for discriminating the results of the analysis in the epidemiology domain. In particular, the representation of the clusters is overlapped to the geospatial distribution of cancer diseases in order to find spatial correlations between cancer incidences and polluted areas. Up to now, no impact of waste treatment on human health has been scientifically proven, but it has not even excluded yet.

A. Introduction

During the last decades the Provinces of Naples and Caserta of Campania region experienced a dreadful increase in the pollution level as effect of documented practices of illegal waste dumping and burning. In the same period, an abnormal increase in deaths from cancer diseases were registered [3].

In order to determine the critical municipalities from the point of view of the urban pollution due to waste disposal sites, we applied a clustering algorithm to assign to each town a hazard index. Furthermore, in order to assign a scale of dangerousness, the index determined was compared with synthetic indicator of municipal risk (IR) that is calculated by (1) that is a domain formula [8]:

 ISBN: 978-1-61208-358-2

$$IR = \sum_{i=1}^{n} S_i * IPP_i * E_i \quad (1)$$

where:

i = number of impact areas in the municipality;

S = surface area that a particular type of waste dump occupies on the municipal territory;

IPP = index of potential hazard

E = index exhibition, it coincides with the resident population involved

B. Data Source and Kinds of Data

The territory of provinces of Naples and Caserta in Campania region (Southern Italy), consisting of 196 municipalities, has got about 300 legal and illegal waste dumping. A part of this area (77 municipalities) has been DeRILID1[3] Vi1Rf1QatARQa31iQt4DM1IRI1UP eDIDARQ' 1E\ 1tKe Italian Ministry of Environment.

TABLE I. DATA ATTRIBUTES

	Attribute	Meaning
1	COM	Municipality name
2	ID_VAL	Socioeconomic deprivation index
3	AREA_IMP_PERC	Surface percentage impacted by dumps
4	POP_AREA_IMP_PERC	Population percentage in surface impacted by dumps
5	4A	Number of the dumps with highest danger level: submerged waste
6	3B	Toxic and hazardous waste, Heaps of dangerous waste
7	2B	Heaps in the pit with the presence of dangerous waste
8	2C	Special waste
9	1D	Storage facilities for non-hazardous waste
10	1E	huge heaps of non-hazardous waste
11	1F	Number of the dumps with lowest danger level: industrial waste
12	TOT_SITI	Total number of dumps

Data came out from [8] where each waste dumping has two indicators: magnitude of the dump and a factor related to the intrinsic hazard of the waste. TABLE I. shows the data attributes used for cluster analysis.

C. Cluster Analysis

The clustering algorithm (*k-means* with k=5) was used to group municipalities.

The result has been chosen according to the purity of the clustering, as well as the geographic representations (Figure 3). Each obtained cluster corresponds to a pollution level associated to the number and types of landfills in the municipal area. The spatial visualization on geographic map by using GeoTagView supports domain experts. In particular, the different colors used to identify clusters on geographical map helps experts to assign a hierarchy level to each cluster and then a level of pollution to each municipality.

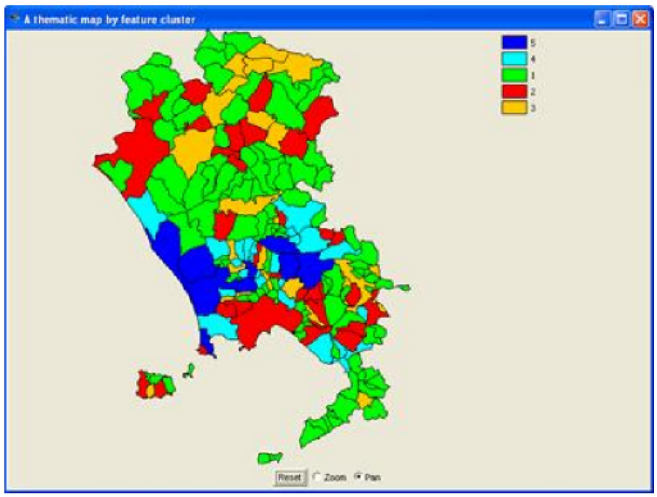


Figure 3. Example of thematic map generated through Weka

Cancer incidences are calculated for each cluster (not in this paper reported).

VII. CONCLUSIONS AND FUTURE WORKS

In this work, we have focused on the problem of how to place emphasis on cluster analysis of data referencing some urban districts. Visualizing such kind of clusters on a geographic map seems to be the more obvious choice in order to show the non-quantitative surface distribution of the feature under examination. So, we extended one of the most used toolkit for the statistical analysis in order to make it able to show a chorochromatic map.

This addition is simple, but it could be very useful in several contexts. We can also make it better, for example by eliminating the preprocessing phase by using external software for building the geospatial database, or by modifying the map view adding information about cluster centroid, and even the medoids indication.

In the end, we could have more extensions in a very simple way, thanks to the great flexibility of the open source toolkit Weka.

ACKNOWLEDGMENT

This work has been carried on within the research project I.D.E.S. ± Intelligent Data Extraction System, funded by the Campania Region and UE within the framework of POR Campania FESR 2007-2013.

REFERENCES

[1] Bogorny V., Palma A.T., “Extending the Weka Data Mining Toolkit to support Geographic Data Preprocessing”. Instituto de Informatica - UFRGS, Porto Alegre, Technical Report – RP-354, 2006.

[2] Bogorny V., Palma A.T., Engel P.M., Alvares L.O., “Weka-GDPM: Integrating Classical Data Mining Toolkit to Geographic Information Systems”. Instituto de Informatica - UFRGS, Porto Alegre, 2006.

 ISBN: 978-1-61208-358-2

[3] Cembalo A. et all., “SOLAP4epidemiologist: A Spatial Data Warehousing Application in Epidemiology Domain”. DaWaK 2013, LNCS 8057, pp. 97-109, Springer-Verlag Berlin Heidelberg 2013.

[1] L. Fazzo et all., “Ecological studies of cancer incidence in an area interested by dumping waste sites in Campania (Italy)”. ANN Ist Super Sanità, Vol.47, No2: 181-191, DOI: 10.4415/ANN_11_02_10.

[4] GeoTools: http://docs.geotools.org/latest/userguide/tutorial/ [retrieved: June, 2014].

[5] D. Guo and L. Mennis, “Spatial data mining and geographic knowledge discovery – An introduction”. Computers, Environment and Urban Systems, 33 (2009), pp. 403-408.

[6] H. J. Miller and J. Ham, “Geographic Data Mining and Knowledge Discovery. An Overview”. In Geographic Data Mining and Knowledge Discovery – Second Edition. Chapman & Hall/CRC, 2009.

[7] Study on the health impact of waste treatment in Campania region (Italy) (2007) http://www.protezionecivile.gov.it/cms/view.php?cms_pk=16909 &dir_pk=395 [retrieved: June, 2014].

[8] I. Turton, “ GeoTools”, in “Open Source Approaches in Spatial Data Handling”, AGIS2, Springer-Verlag, Berlin, Heidelberg, 2008.

[9] H. I. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”. Elsevier – Morgan Kaufmann, 2005.

[10] Projects related ɒ Weka http://weka.wikispaces.com/Related+Projects [retrieved: June, 2014].

[11] Kea – Keyphrase extraction algorithm http://www.nzdl.org/Kea/ [retrieved: June, 2014].

 ISBN: 978-1-61208-358-2

Document Identification with MapReduce Framework

Yenumula B Reddy
Department of Computer Science
Grambling State University,
Grambling, USA
email: ybreddy@gram.edu

***Abstract*—Hadoop technology made a break through to process the unformatted data and generates the results faster than ever. Before Hadoop technology, the results were produced for formatted data using SQL and other techniques. They allowed to effectively sharing memory, central processing unit, disk, and network input/output more efficiently. There was no proper system to analyze unformatted data. The paper discusses the MapReduce framework to identify a required document from a stream of documents. We proposed an algorithm called MapReduce to detect sensitive documents that identify sensitive or required document among the streams of documents. The algorithm was tested using the Hadoop package and Java program. The results conclude that the Java program is useful for small documents. The Hadoop technology helps in stream of documents and produces the results much faster than simple Java program implementation.**

Keywords-MapReduce; Hadoop Distributed File Systems; Big data; key, shuffle; Apache Zookeeper.

I. INTRODUCTION

Big data (lower case is used for *Big data* in most of the places unless it is required to stress the word *Big*) is a general term used for a large volume of data that are structured, unstructured and semi-structured data created or generated by a company. This data cannot be loaded using any database models and it is not possible to get results with available query languages. That means that this data cannot be processed using traditional tools. The data is not a specific quantity in terms of a number of bytes (terra-bytes, petabytes or exabyte). It is continuously growing in size every minute. It is real big and grows in Exabyte. The origins are from a variety of platforms. That data volume changes quickly and its growth cannot be predicted. It is expensive to analyze, organize, and preparing as useful data. Therefore, we need special effort to prepare as meaningful by using new algorithms, special hardware and software. There may not be a single tool to analyze and transform big data as meaningful data. The tools may vary to analyze data. The big data management needs a generating high quality of data to answer the business queries.

The primary goal is to discover the repeatable business patterns, uncover the hidden patterns, and understand the unknown correlations and other useful information. The business giants have access to the information, but they do not know to get the value out of this unstructured data. The traditional tools are not helpful due semi-structured storage. The big data analytics may commonly use software tools such as predictive analytics [16] and data mining [17]. The technology associated with big data analytics includes NoSQL databases [21], Hadoop [2] and MapReduce [4]. These are open source frameworks that support processing of big data. The analysis needs MapReduce to distribute the work among a large number of computers and process in real-time. Hadoop structure lowers the risk of catastrophic system failure, and even when a significant number of nodes become inoperative.

The government and information technology managers are highly motivated to turn the massive data into use. Today, Hadoop framework and MapReduce, offer new technology to process and transform Big data into meaningful data. It is required to deploy the infrastructure differently to support the distributed processing and meet the real-time demands. The Information Technology (IT) managers found that security and privacy are the major problems, particularly while using third-party cloud service providers [15].

The structured and unstructured data come from a variety of sources. Adoption of big data tools to process the big data is increasing. High priority is given to improving the big data formalizing and processing. The top data for transactions include business data documents, email, sensor data, image data, Weblogs, Internet search indexing, and attached files. The IT group found an interest in learning about technology, deploy the packages to process the data, and adopt the infrastructure for better performance and affordable cost. They are in the process of implementing Apache Hadoop frameworks, and commercial distributions of the Hadoop distributed framework and other NoSQL databases.

Currently, priority is given to processing and analyzing unstructured data resources including Weblogs, social media, e-mail, photos, and videos. Unstructured emails are given priority to analyze and process. As a first step, the IT staff is working on batch processing and move to real-time processing. The companies are concerned about the scalability, low latency, and performance in storing and analyzing the data. Further, they are worried about protection of data for third-party cloud providers. Standards are required for data privacy, security, and interoperability for data and systems.

 ISBN: 978-1-61208-358-2

The big data are useful when analyze and store in a meaningful way, so that the data can be accessed immediately. The main goal is to store the data to ensure that the data are accessible for business intelligence and deepanalysis. It is required to design a tool to analyze the data and provide the answers with minimum efforts and time. The challenges include the size, structure, origin, variety, and continuous change in data. The data are real big in size (terra-bytes or exa-bytes) and unstructured. Data contains text, figures, videos, tweets, Facebook posts, website clicks, and different types of data from a variety of websites. The origins are varied and come from a variety of platforms and multiple touch points. Data change fast in terms of format, size, and types of websites. Further, specialized software is required to pull, sort, and make the data meaningful. There is no universal solution to make such data meaningful. Further, the data are produced in universally available languages. Processing such data may need separate algorithms (depending upon the data). There are many predictions in years to come about the data coming from an unknown source and unstructured in nature. The predictions include the standardization and marketing. The following predictions include the data origin, type, size and management [23][24].

- Massive data collection across multiple points
- Firmer grip on the data collected by different groups
- Generalization of format for Internet data and/or data generated by business media
- Entering of social media as part of big data generation

The main purpose of the data management (analyzing and processing) is to make sense of the collected data from various data collection points. Making sense of data means that the end point of the processing of data must be able to answer the business queries. The queries include data mining related predictions, business queries, and management assistance.

The Hadoop project was adopted more in the Department of Defense than in other agencies. These agencies could not use the Hadoop system design because the Hadoop design lacks reusability due to Java Application Programming Interfaces for data access in Apache. Research is required in understanding the Hadoop system design, security models, and usage in a specific application. Protocol level modifications help in improving the security at source level.

Federated systems enable collaboration of various networks, systems and organizations at different trust levels. Clients must be separated from service with authentication and authorization procedures. Existing security models protect the resources within the boundary of the organization. In federated systems, new participants join and leave continuously. They may not all be trusted. Therefore, federated systems require the security specification for each function. Depending upon the system sensitivity level the boundary constraints are incorporated. Individual protection domains are required for each entity. Therefore, the existing security domain procedures do not work in federated systems.

The federated systems are distributed. The security system in federated systems separates the client access, authentication, and authorization. Therefore, they need collaboration between networking, companies and associated systems. Due to the involvement of many entities in the federated systems, it is difficult to maintain the security among these entities. The threat may be expected (unavoidable) from a variety of sources. Hence, a high-level coarse-grained security goals need to be specified in the requirements.

In this paper, we discuss the Hadoop single node installation and introduced an algorithm to identify the sensitive documents. Section 2 discusses the current state of Hadoop distributed file system. MapReduce programming model is discussed in Section 3 and implementation in Section 4. Finally, Section 5 discusses the conclusions and future work.

II. HADOOP DISTRIBUTED FILE SYSTEMS – CURRENT STATUS

Data backup in Hadoop file systems using snapshots was discussed by Agarwal et al. [1]. The authors designed an algorithm for selective copy on appends and low memory overheads. In this work, the snapshot tree and garbage collection was managed by Name-node. The architecture of Hadoop Distributed File Systems (HDFS) and the experiences to manage 25 petabytes of Yahoo data was discussed by Shvachko et al. [2]. The fault tolerant google file system running on inexpensive commodity hardware with aggregate performance to a large number of clients was discussed by Ghemawat et al. [3]. The paper discusses many aspects of design and provides a report of measurements from both micro-benchmarks and real world use. The MapReduce programming model was explained and it's easy to use functions were discussed by Dean and Ghemawat [4]. The document discussed the experiences and lessons learned in implanting the model. Further, it discusses the impact of slow machines in redundant execution, locality of optimization, writing single copy of the immediate data to local disc, and saving the network bandwidth. Chang et al. [5] discussed the dynamic data control over data layout and format using 'Bitable' a flexible solution. They claimed that many projects are successfully using this model by adding more machines for process over the time.

The federated security architecture was discussed in Windows Communication Foundation (WCF) [6]. WCF incorporates the security in federated systems to build and deploy the federated systems. The architecture of these federated systems includes the federation, domain/realm, security token service, and consists of primary security architecture. The current mobile devices were implemented with imitation of 60K tasks, but the next generation of Apache MapREduce supports 100K concurrent tasks [7]. In MapReduce, users specify the computation in terms of the map, and a reduce function. The available software modules in MapReduce automatically paralyze the computation and schedule the parallel operations with the help of network and computational facilities. These facilities help to complete the operation much faster. Every day, an average of a hundred

 ISBN: 978-1-61208-358-2

thousand MapReduce jobs are executed on Google clusters [4]-[7].

Halevy et al. [8] discussed that a nonparametric model is needed to represent the data in large data sources. They believe that a nonparametric model holds a lot of details compared to a parametric model. The authors believe that the selection of unsupervised learning on unlabeled data generates better results than on labeled data. Halevy et al. [8] pointed out that future research includes the creation of specific data sets by automatically combining data from multiple tables. Combing data from multiple tables also includes unstructured Web pages or Web search queries. Thuraisingham [9] discussed various types of security policies including local, generic, component, export and federate. The security policy generation enforcing also included in this study.

Hadoop security was discussed in the reports [10]-[15]. Reddy [10] proposed the security model for Hadoop systems at access control and security level changes depending upon the sensitivity of the data. Authentication and encryption are the two Security levels for big data in Hadoop. Ravi [11] concludes that Kerberos files keep the intruders away from accessing the file system and Kerberos system has better protection compared to other federated systems. Chary et al. [12] discussed the current level of security in Hadoop distributed file systems that include the client access for Hadoop cluster using Kerberos Protocol and authorization to access.

O'Malley et al. [13] and Das et al. [14] discussed the security threats from different user groups in Kerberos authentication system. The research in Kerberos and MapReduce implementation details also discusses the security, role of delegation token. The research in [13] and [14] emphasizes the need for security in internal and external access level for Hadoop systems. The work describes the need for limits of access rights to specific users by application, isolation between customers, information protection and incorporation of encryption models.

Srinath [15] presented "Airavat", a MapReduce-based prototype system, provides strong security, privacy and guarantees for distributed computations of sensitive data. The model described in "Airavat" explains how to use different parameters, estimate their values, and test on several different problems. This system does not follow the software engineering methodology. Therefore, it has weak use cases and complicated processes in specifying the parameters. Since MapReduce computations are not efficient, organizations raised critical questions on privacy and trust of data during the MapReduce computations.

Preserving the privacy in big data was discussed by McSherry [16]. McSherry's Privacy integrated queries (PINQ) presents an opportunity to establish a more formal and transparent basis for privacy technology. The algorithms designed help the users in increasing the privacy-preserving and increases the portability. Partha et al. [17] presented a system that learns for data-integrated queries which use sequences of associations. The associations include foreign keys, links, schema, mappings, synonyms, and taxonomies. They create multiple ranked queries linking the matches to keywords. These queries are linked to Web-forms and users have only to fill the keywords to get answers. MapReduce application was used for integer factorization [18], Matrix computation [19], and machine learning on multicore systems [20]. None of these papers discussed the detection of sensitive data files among the streams of data files. This paper introduces an algorithm called **M**apReduce **A**lgorithm to **D**etect **SE**nsitive **D**ocuments (MADSED).

III. MAPREDUCE PROGRAMMING MODEL

One of the programming models in MapReduce is breaking the large problem into several smaller problems. The solutions to the problems are then combined using the MapReduce function to give a solution to the original problem. The functionality is similar to software engineering top-down design. There are many questions that arise in MapReduce application due to dynamic input, nodes may fail, number of smaller problems may exceed the number of nodes, dependable sub-problems, distribution of input to the smaller jobs, coordination of nodes, and synchronization of completed work. MapReduce programming model and the associated implementation can be used to solve these problems by processing and generating large data sets.

MapReduce application divides into three parts: Map, Shuffle and Sort, and Reduce. A Map part of MapReduce job splits the input data-set into independent chunks. The independent chunks are processed in a completely parallel manner using Map task. A given input pair can have zero or more output pairs. The map-outputs are merged and constructed with respect to the key values. These pairs are propagated to reduce function. The reduced function then merges these values to form a possibly smaller set of values. That is, the reduced function filters the map output and produces the results with respect to key. The total functionality includes scheduling the tasks, monitoring them, and re-executes the failed tasks.

The mapper breaks down the problem into smaller bits. These bits are processed parallel to produce a solution. The formula (1) produces new key values.

$$\forall(k,v)\rightarrow(k^{'},v^{'}) \quad (1)$$

where, k = key, v = value. These values are used for searching the keywords in another semantic domain to produce intermediate values (document identification and document) for each call. In Figure 1a, A is document identification and α is a document. The mapper then combines related keys and prepare the partitions to search in the documents. The map-outputs are sorted, merged and constructed with respect to the key values in the shuffle step. The aggregated values are filtered and then return a new set of results in the reduce phase. After completion of reducing phase, it returns all possible values with respect to a key. The process is shown in Figures 1a [21] and 1b [22].

For example, group the words of same length in the statement "*The Big data is useful if we analyze and stored in a meaningful way so that the data can be accessed quickly*". The MapReduce framework groups all of the values by key

 ISBN: 978-1-61208-358-2

(each word). The Table I shows the output of key (the number of letters in a key and key word) called value pairs. The same procedure will be applied for each document (one document or multiple documents). Each document will be treated separately at the time of process in MapReduce function.

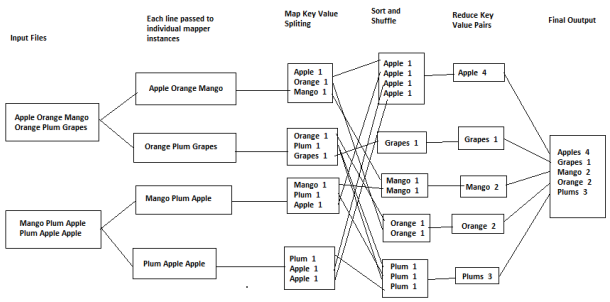


Figure 1a. Map, Shuffle, and Reduce phase example.

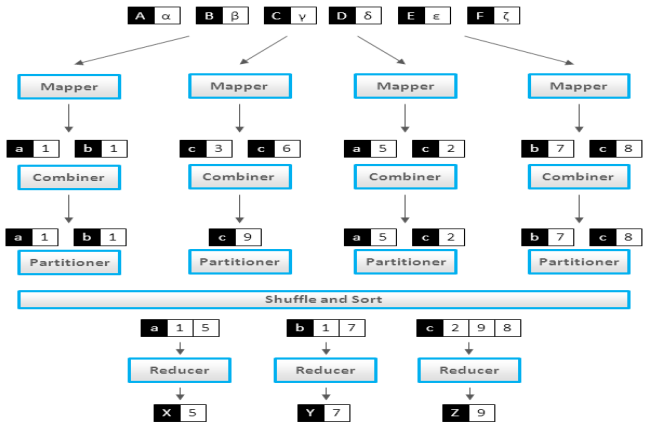


Figure 1b. Map, Shuffle, and Reduce phase example.

TABLE I. KEY AND WORD

3: The	7: analyze	2: so
3: big	3: and	4: that
4: data	6: stored	3: the
2: is	2: in	4: data
6: useful	1: a	3: can
2: if	10: meaningful	2: be
2: we	3: way	8: accessed
		7: quickly

In Table I, the keys are grouped according to the number of letters in a word (1 letter word or 2 letter word, etc.) and figure out number of items in each key. The number of items in each key (1 letter word or 2 letter word etc.) is shown in Table II. The reduce function counts the number of items with key size of the list (Table III). The reduction can be done in parallel using Graphics Processing Units (GPUs).

TABLE II. WORDS RELATED TO EACH KEY

1: a
2: is, if, we, in, so, be
3: the, big, , way, the, can
4: data, and, that, data
6: useful, stored
7: analyze, quickly
8: accessed
10: meaningful

TABLE III. KEY WITH THE SIZE APPEARS IN THE LIST

1: 1
2: 6
3: 5
4: 4
6: 2
7: 2
8: 1
10: 1

The algorithms for Map, Shuffle, and reduce are given below. The MapReduce algorithm to generate Table I, II and III for all documents has three phases namely, mapper, combiner, and reducer.

```
class Mapper
  method Map (Document-id id, document d)
      for each term t in document d
            store the term and its size

class Combiner
   method combine (term t, [c1, c2, …])
      sum =0
      for each term t in list [c1, c2, …]
            append to list of same size terms
      complete for different size terms

class Reducer
  method Reduce (term t, counts [c1, c2, ….])
    for count c in [c1, c2, …]
    count the same size terms and store number of
    occurrences of each term in the document
repeat this for all terms
```

The algorithm generates the tables similar to Table I, II, an III by using this algorithm from the given document or documents. This algorithm is enough to find the number of occurrences of each term. We need the extension of the algorithm to detect the importance or sensitivity of the

document. The algorithm for detecting the required documents among the documents with the help of keys and their weights is called MapReduce Algorithm to Detect Sensitive Documents (MADSED).

MADSED Algorithm

Let us assume that the key sizes are: 3, 5, 6, 7 letters

- Divide the document into N sub-documents
- Filter each sub-document by leaving only words of sizes 3, 5, 6, and 7
- Count # of times each key word appears in the sub document
- Shuffle the sub document results into single output with number of times the keyword appears
- Calculate value of each key word by multiply each keyword by its weight and times appear
- Add each keyword output values as result
- If the result (output) is greater than threshold value, the document is important; if it is boarder on threshold, it is for consideration; otherwise reject
- End of algorithm

The Algorithm was implemented using Hadoop technology and Java program implementation.

IV. IMPLEMENTATION

The Hadoop 1.2.1 from the Apache Website was installed on Linux operating system [25]. The JavaTM 1.7.x was installed as part of the installation [26]. Initially, we tested a single node installation. It has login and password protection. We used a text file from the Hadoop folder by using the commend (2).

```
bin/hadoop dfs -copyFromLocal
/home/csadmin/Downloads/gutenberg /user/csadmin/Gutenberg.
```
(2)

The commands and their usage are available at http://hadoop.apache.org/docs/r2.3.0/hadoop-project-dist/hadoop-common/FileSystemShell.html

For more information http://www.apache.org/

HDFS as a general DFS for applications are available at http://www.opensourceforu.com/2013/12/peek-hadoop-distributed-file-system/

To make sure the file is copied successfully use the following command

```
bin/hadoop dfs -ls /user/csadmin
```

The command to run the MapReduce is

```
bin/hadoop jar hadoop-examples-1.2.1.jar wordcount
/user/csadmin/gutenberg /user/csadmin/gutenberg-output
```

Finally, the output generated by MapReduce is checked. It generates all words and number of times each word repeated. Table IV shows the sample output of first few words.

TABLE IV: SAMPLE OUTPUT OF MAPREDUCE

"(Lo)cra"	1
"1490	1
"1498,"	1
"35"	1
"40,"	1
"A	2
"AS-IS".	1
"A_	1
"Absoluti	1
"Alack!	1
"Alack!"	1
"Alla	1

As a next step, we used the keywords and stored only those keywords and number of times each keyword repeated from the generated output. Then, we multiplied each keyword with its weight and number of times repeated and adds all the resulted values. For example, the keywords are Oil and Alaska. If the weight for oil = 0.02 and Alaska is 0.1, the oil repeats 15 times in the document and Alaska repeats 20 times in the document. The total value is 0.02*15 + 0.1*20= 0.3+2.0=2.3. If the total value greater than the threshold value, then the document is retrieved (*the threshold value is established by the programmer depending upon the expected minimum value required to select the document*). The algorithm was implemented using Java program and generated the results. The results produced through Java program in the current case (to select a required document) are compatible (means document is selected or rejected) to Hadoop results. The Java program is good for small files and takes more time in case of large files. Hadoop technology produces the results much faster and recommended for large files. The algorithm is useful for detecting the sensitive data or files whenever a stream of files is entering in the organization. The algorithm separates the sensitive files from a large set of files. Further, the process helps to retrieve required files among thousands of files.

V. CONCLUSIONS AND FUTURE RESEARCH

The paper discussed the MapReduce framework to identify an essential document from a stream of documents. The research discusses the current state of Hadoop distributed File Systems, MapReduce programming model, and presents an algorithm to identify sensitive or required document among the streams of documents. The algorithm was tested using the Hadoop package and Java program. The results conclude that the Java program is useful for small documents, whereas Hadoop helps in large and a stream of documents. Further, Hadoop produces the results must faster than simple Java program implementation.

The future research involves testing the documents with Hadoop multimode (installed and tested) and Hadoop with Graphics Processing Unit (GPU) based technology. The work is underway and results will be available soon. Big data security is another problem that is not discussed in this paper.

 ISBN: 978-1-61208-358-2

ACKNOWLEDGEMENTS

The research work was supported by the AFRL Collaboration Program – Sensors Research, Air Force Contract FA8650-13-C-5800, through subcontract number GRAM 13-S7700-02-C2.

REFERENCES

[1] S. Agarwal, D. Borthakur, and I. Stoica, "Snapshots in Hadoop Distributed File System", UC Berkeley Technical Report UCB/EECS, 2011.

[2] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System", 26th IEEE (MSST2010) Symposium on Massive Storage Systems and Technologies, May, 2010, pp. 1-10.

[3] S. Ghemawat, H. Gobioff, and S. Leung, "The Google File System", SOSP'03, October 19–22, 2003, pp. 29-43.

[4] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Proc. 6th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2004, San Francisco, USA, Dec. 2004, pp. 107-113.

[5] F. Chang et al., "Bigtable : A Distributed Storage System For Structured Data", ACM Transactions on Computer Systems (TOCS), vol. 26, Issue 2, June 2008, pp. 204-218.

[6] Authentication, Microsoft Patterns & Practices, http://msdn.microsoft.com/en-us/, 2012 [Retrieved: July 27, 2014].

[7] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters", CACM 50th anniversary issue, vol. 51, issue 1, Jan 2008, pp. 107-113.

[8] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data", IEEE Intelligent Syst., 2009, pp. 8-12.

[9] B. Thuraisingham, "Security issues for federated database systems", Computers & Security, 13 (1994), pp. 509-525.

[10] Y. B. Reddy, "Access Control for Sensitive Data in Hadoop Distributed File Systems", Third International Conference on Advanced Communications and Computation, INFOCOMP 2013, November 17 - 22, 2013 - Lisbon, Portugal, pp. 72-78.

[11] P. Ravi, "Security for Big data in Hadoop", http://ravistechblog.wordpress.com/tag/Hadoop-security/, April 15, 2013 [Retrieved: July 27, 2014] .

[12] N. Chary, K. M. Siddalinga, and Rahman, "Security Implementation in Hadoop", http://search.iiit.ac.in/cloud [retrieved: January 2013].

[13] O. O'Malle, K. Zhang, S. Radia., R. Marti, and C. Harrell., "Hadoop Security Design", http://techcat.org/wp-content/uploads/2013/04/Hadoop-security-design.pdf, 2009, [Retrieved: July 27, 2014].

[14] D. Das, O. O'Malley, S. Radia, and K. Zhang, "Adding Security to Apache Hadoop", Hortonworks Technical Report 1, http://www.Hortonworks.com, 12 pages.

[15] I. Roy Srinath, T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and Privacy for MapReduce", 7th USENIX conference on Networked systems design and implementation (NSDI'10), 2010, Berkeley, CA, pp. 1-16.

[16] F. McSherry, "Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis", Proceedings of SIGMOD, 2009, pp. 19-30.

[17] P. P. Talukdar et al. "Learning to Create Data-Integrating Queries", VLDB, 2008, pp. 785-796.

[18] J. Tordable, "MapReduce for Integer Factorization", arXiv, arXiv:1001.0421v1 [cs.DC], January 4, 2010. [Online]. http://arxiv.org/abs/1001.0421v1

[19] S. Seo, E. J. Yoon, J. Kim, S. Jim, J. Kim, and S. Maeng, "HAMA: An Efficient Matrix Computation with the MapReduce Framework", IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), 2010, pp. 721-726

[20] C. Chu et al., "Map-Reduce for Machine Learning on Multicore", Advances in Neural Information Processing Systems 19 (NIPS 2006) pp. 281-288.

[21] Highly Scalable Blog, Articles on Big data, NoSQL, and Highly Scalable Software Engineering, MapReduce Patterns, Algorithms, and Use Cases, http://highlyscalable.wordpress.com/, Posted on February 1, 2012 .

[22] K. S. Bejoys, "Word Count - Hadoop Map Reduce Example", http://kickstarthadoop.blogspot.com/, April 29, 2011 [Retrived: July 27, 2014] [Retrieved: July 27, 2014].

[23] E. Namey, G. Guest, L. Thairu, and L. Johnson, "Data Reduction Techniques for Large Qualitative Data Sets", Handbook for team-based qualitative research, pp. 137-161.

[24] J. Karlsrud, "Peacekeeping 4.0: Harnessing the Potential of Big Data, Social Media and Cyber Technology", in Kremer, J. F. & Müller, B. (eds.) Cyber Space and International Relations. Theory, Prospects and Challenges. Berlin: Springer,2013, pp. 141-160.

[25] http://hadoop.apache.org/docs/r1.2.1/ 2011 [Retrieved: July 27, 2014].

[26] http://en.softonic.com/s/java-1.7 [Retrieved: July 27, 2014].

 ISBN: 978-1-61208-358-2

The Economic Benefits of Allocating Spectrum for Mobile Broadband in Korea

An Input-output Anlysis

Jae hyouk Jahng
Industrial Strategy Research Department of ETRI
Daejeon, South Korea
sapaha@etri.re.kr

Abstract— **In South Korea at the year-end 2013, the spectrum regulatory agency finalized the "Mobile Gwanggaeto Plan 2.0" for finding and supplying 1190 MHz bandwidth of spectrum that will be allocated for mobile communication in 4 phases by 2023. So the main purpose of this study is to find an economic impact on mobile broadband spectrum allocation below of 3.6 GHz with economic benefits by means of estimating service revenues and using an input-output analysis. The newly-added benefits will be 159.6 trillion Won and creating more than 182,500 jobs over 7 years.**

Keywords- spectrum; input-ouput analysis; economic benefits.

I. INTRODUCTION

The global market scale of mobile communications has manifested a trend of continuous while the development of the third-generation (3G) network evolving into Long Term Evolution (LTE) has accelerated its pace. The Global mobile Suppliers Association (GSA) report says that 300 LTE networks around the world have been put into commercial service in 107 countries up to June 2014, and forecasts 350+ commercially launched LTE networks by end 2014 [1].

Currently, South Korea's LTE service has grown faster than expected. For instance, LTE penetration rate to exceed 58% by May 2014 [2], and LTE traffic rates on the network is 89% in comparison with mobile communication [3].

South Korea has assigned 390 megahertz (MHz) for Mobile systems so far, as shown in Table I. 330 MHz of the spectrum is utilized to the frequency division duplex (FDD), else 60 MHz to the time division duplex (TDD) as mobile WiMAX.

TABLE I. SPECTRUM ALLOCATION FOR MOBILE IN SOUTH KOREA

Band (MHz)		Uplink		Band width	Downlink		Sum-up
		lower	*upper*		*lower*	*upper*	
2G	800	824	829	5	869	874	10
	1800	1770	1780	10	1860	1870	20
3G	2100	1940	1980	40	2130	2170	80
WiMAX	2300	TDD		60	2300	2360	60
LTE	800	819	824	5	864	869	50
		829	849	20	874	894	
	900	905	915	10	950	960	20
	1800	1715 1730	1725 1735	15/20	1810	1830	70
		1735 1745	1740 1755	15/20	1830	1850	
	2100	1920	1940	20	2110	2130	40
	2600	2520	2540	20	2640	2660	40
Sum-up							390

But according to the Ministry of Science, ICT & Future Planning (MSIP) spectrum regulatory agency in Korea, South Korea wireless networks carried approximately 444 petabytes per month in 2023 [4], a greater than 26-fold increase in comparison with 2011.

The growth of wireless broadband will be constrained if government does not make spectrum available to enable network expansion and technology upgrades. In the absence of sufficient spectrum, network providers must turn to costly alternatives, such as cell splitting, often with diminishing returns. And also the progression to 4G LTE technologies may require appropriately sized bands, including larger blocks to accommodate wider channel sizes.

In order to meet growing demand for wireless broadband services, and to ensure that MSIP finalized "Mobile Gwang-gae-to Plan 2.0" for finding and supplying 1190 MHz bandwidth of spectrum will be allocated for mobile communication like LTE in 4 phases by 2023 [4]. The term "Gwang-gae-to" is the most famous king name for the territorial expansion in Korean history and that means "broadband". 110 MHz bandwidth has already secured, including 40 MHz bandwidth in the 700 MHz band, 30 MHz bandwidth in the 1.8 GHz band and 40 MHz bandwidth in the 2.6 GHz band. Of this total amount of 1190 MHz bandwidth between 700 MHz and around 6 GHz should be made available for mobile use within 10 years. The timeline in Table II illustrates a schedule of actions. This plan had to be modified in consideration of the accelerated increase of mobile traffic due to the evolution of mobile communication technologies, expiration of the use of existing bands and increasing demands for broadband spectrums.

TABLE II. THE MOBILE GWANGGAETO PLAN 2.0 IN KOREA

Category (MHz)		Secure Possible Bandwidth (MHz)					Total
		Secured	*Phase1*	*Phase2*	*Phase3*	*Phase4*	
FDD	700	40					40
	1800	30			20(20)		50(20)
	2100	-	(60)	60			60(60)
	2600	40	20			10	70
TDD	2000			40			40
	2300			30(40)			30(40)
	2500		40				40
FDD/TDD	3500			160			160
WRC	6000				200	500	700
Total	Adding	110	60	290	220	510	1190
	Refarming		(60)	(40)	(20)		(120)

 ISBN: 978-1-61208-358-2

The main purpose of this study is to find an economic impact on mobile broadband spectrum allocation with economic benefits by means of estimating service revenues and using an input-output analysis.

The rest of this paper is organized as follows. Section II describes the study design issue. Section III goes into finer details results with respect to the contribution of Gross Domestic Product (GDP) and job creation. Section IV addresses the conclusion remarks.

II. STUDY DESIGN ISSUE

A. The past research and literature

Both historical and international experiences have testified that the mobile Internet can contribute to economic growth. Of course, this is not just due to the need of production input to build networks and sell mobile phones, but more importantly, due to the fact that the mobile Internet can advance the spread of information, improve productivity and efficiency, and enable individuals to explore new market and services in the whole economy. Some of the studies highlighted below in this key source suggest that countries can accelerate economic growth and productivity by increasing mobile broadband Internet adoption and usage.

Deloitte has estimated that a 10% rise in 3G penetration increases GDP per capita growth by 0.15 percentage points [5]. Plum report demonstrates that mobile services generate the greatest economic value of eight major applications of spectrum, and estimates the economic value of spectrum used for mobile could reach €477 billion by 2023 [6]. Other report estimates that the reassignment of 300 MHz of spectrum to mobile broadband within five years will spur $75 billion in new capital spending, creating more than 300,000 jobs and $230 billion in additional GDP [7].

A joint research by the GSM Association (GSMA) and the Boston Consulting Group (BCG) has also found that if Asia-Pacific nations were to allocate the 108 MHz spectrum (698 to 806 MHz) to mobile broadband, then in the period from 2014 to 2020, this would contribute 68.4 billion USD to the GDP growth in South Korea [8].

B. Loglet Analysis

Many diffusion models outline the consumption styles of homogeneous consumers in the same population. However, a specific service can be divided into a few heterogeneous groups, showing different diffusion processes. For instance, the mobile service is one communication service with several different generation subscribers – 2G, 3G and 4G. According to related subscribers and service income of mobile communications network, we have chosen the S-shaped analysis and prediction of complex growth processes as the basic formula for our forecast. In this study, the Loglet analysis model will be used, which provides the tool for analyzing the demand diffusion patterns of mobile broadband groups separated from the entire mobile communications, using the business plan of mobile service operators and the series of the demand-data collected so far. Particularly, this study aims primarily to make a mobile broadband-use demand diffusion patterns, which shows the fast growth in demand along with 4G LTE service in Korea.

The Loglet is composed with various logistic curves. The solution to the logistic differential (1) is :

$$P(t) = \frac{\kappa}{1 + \exp(-\alpha(t - \beta)} \quad (1)$$

(1) has three parameters; the saturation point of demand **κ**, the growth time reaching a 90% saturation point from the 10% one **α**, and the turning point of the logistic curve **β**. The parameter **α** is usually more useful than **△t** for the analysis of historical time-series data. The parameter **β** specifies the time when the curve reaches 1/2 **κ**, often re-labeled **tm.**

The formula below illustrates the Loglet as the sum of logistic models. Through simple algebra, the characteristic duration is related to **α** by ln(81) /**△t**. The parameter **β** specifies the midpoint of the growth trajectory as **tm**. The parameter **κ** is the asymptotic limit that the growth curve approaches as like market niche or carrying capacity [9].

The formula below in (2) illustrates the Loglet model for the total mobile broadband subscribers as the sum of logistic models. The three parameters **κ**, **△t**, and **tm** define the parameterization of the logistic model used as the basic building block.

$$N(t) = \frac{\kappa}{1 + \exp\left[-\frac{\ln(81)}{\Delta t}(t - t_m)\right]} \quad (2)$$

C. The input-output model

The input-output (I-O) model is the most commonly applied method for economic analysis. Its developer, Wassily Leontief, received the Nobel Prize for Economic Science in 1973 [10]. Within the I-O framework, transactions between industries are modelled to examine the quantitative relationship between the supply of industry inputs and the range of produced outputs.

The I-O model requires that the economy be divided into sectors. Each sector produces goods or services except for the final sector, which only consumes goods and services. A production vector X lists the output of each sector. A final demand vector D lists the values of the goods and services demanded on other productive sectors by the final sector. These intermediate demands are described by the consumption matrix A for the economy.

$$A_{ij} = \frac{Q_{ij}}{Q_i} \quad (3)$$

The equilibrium levels of production for each sector may now be calculated. These equilibrium levels are the production levels which will just meet the intermediate

 ISBN: 978-1-61208-358-2

demands of the sectors of the economy plus the final demands of each sector. If X is the desired production vector, X must satisfy

$$X = AX + D \quad (4)$$

In (4), X represents a matrix of gross sector output, A represents a matrix of technical coefficients yielding the input required by one sector to produce a unit increase in a linked sector, and D represents a vector of final demands.

This study used noncompetitive imports and domestic the latest I-O tables for 2009 distributed by Korea's central bank, the Bank of Korea (BOK) to examine the inter-sector effects [11]. So the demand-driven model is used to investigate the production-inducing effect of the mobile broadband service sector on the economy using an exogenous specification. The effect is calculated solving the gross output necessary in each sector for a given set of final demands yields.

$$X = (I - A)^{-1} D \quad (5)$$

In (5), $(I - A)^{-1}$ represents the Leontief inverse matrix or input inverse matrix indicating the total effects on production of a unit increase in final demand with I as the identity matrix. In other words, Column sums of elements in the Leontief inverse describe how an increase in the final demand of a sector impacts on the production levels of all other sectors in the economy and is referred to as the backward linkage effect. Backward linkages provide a measure of each sectors' relative importance as a demander from other upstream suppliers.

Table Ⅲ show the final demand multipliers for the "mobile broadband services" industry that can be used to estimate the impacts of the increase of mobile service revenue on the domestic economy. The total impacts that are calculated using multiplier from this table will include the final-demand change, as well as the direct, indirect effects. This suggests that an increase of $1 in mobile broadband services results in an increase of $1.6864 in final Korea output (or GDP). Using the same hybrid approach, the estimated jobs multiplier for mobile broadband is 7.7007. That is, an increase of $1 million (almost 1 billion Won) in mobile broadband results in roughly 7.7 new Korea jobs.

The increase in mobile broadband revenue may be broken down into the impacts on other industrial sectors in the national level. Therefore, this study is predicting sectoral linkage effects of mobile broadband services, indirect effect is the sum of them by using exogenous specifications for the mobile broadband sector in I-O tables.

TABLE III. FINAL-DEMAND MULTIPLIERS OF MOBILE SERVICES

Industry	Output *(inducement coefficient)*	Jobs *(employ inducement coefficient)*
direct	1.0000	2.8259
indirect	0.6864	4.8748
Sum-up	1.6864	7.7007

III. RESULT OF ANALYSIS

In order to clarify the feasibility and accuracy of this study, the hypothesis involved include a few aspects as follows; Firstly, we classify scenario into the circumstances additional spectrum by increments of 700 MHz below of 3.6 GHz band. Secondly, spectrum amount of 60 MHz mobile WiMAX is excluded in the economic effect, because WiMAX and LTE are strictly divided into different market in Korea government policy. Thirdly, existing use of mobile communication spectrum is 330 MHz. Therefore, analyzing scenario allocate an additional 490 MHz to mobile broadband giving a total of 820 MHz.

On the basis of related parameter estimated by the growth trend in the past years, we have forecasted the mobile broadband subscribers by using the subscribers forecast from the Loglet in Table Ⅳ. And LTE service revenue of mobile communications in the seven-year period from 2014 to 2020 by using the mix of subscribers and year-on-year ARPU in Table Ⅴ. And also we have assumed that the average revenue per user (ARPU) growth will be 3% every step of large supply of frequency by appearance of innovative services in light of annual growth rate of GDP, after then applied to the ARPU decreasing pattern.

The above scenarios, based on the input-output model, the economic impact of spectrum is shown in the following Tables Ⅵ, Ⅶ, Ⅷ. The economic effects of mobile broadband service will be 40 trillion Won, and job creation will be 182.5 thousand man-year in the year 2020. Using an inducement coefficient of 1.6864 and employ inducement coefficient 7.7007.

TABLE IV. MOBILE BROADBAND SERRVICE FORECAST

	2014	2015	2016	2017	2018	2019	2020
Rate (%)	19.2	29.0	41.0	54.0	66.0	75.8	82.9
Subs.(000)	10,985	16,959	24,511	32,856	40,855	47,589	52,715

TABLE V. REVUNUE OF MOBILE BROADBAND SERVICE (TRILLION WON)

	2014	2015	2016	2017	2018	2019	2020
Revenue	4.8	7.6	11.0	14.6	18.4	21.6	23.7

TABLE VI. ECONOMIC EFFECTS OF MOBILE SERVICE (TRILLION WON)

	2014	2015	2016	2017	2018	2019	2020
Direct	4.8	7.6	11.0	14.6	18.4	21.6	23.7
Indirect	3.3	5.2	7.6	10.0	12.6	14.8	16.3
Total	8.1	12.8	18.6	24.6	31.0	36.4	40.0

TABLE VII. EMPLOY EFFECTS OF MOBILE SERVICE (THOUSAND PERSON)

	2014	2015	2016	2017	2018	2019	2020
Direct	13.6	21.5	31.1	41.3	52.0	61.0	67.0
Indirect	23.4	37.0	53.6	71.2	89.7	105.3	115.5
Total	37.0	58.5	84.7	112.5	141.7	166.3	182.5

TABLE VIII. ECONOMIC BENEFITS OF SPECTRUM DEPENDENT SERVICES

Sectors	**Y2020** *(Adding 820 MHz)*
Economic effects of mobile service	40.0 *(Trillion Won)*
Job Creation	182.5 *(Thousand man-year)*

 ISBN: 978-1-61208-358-2

IV. CONCLUDING REMARKS

Spectrum has a prominent contribution to economic growth, albeit with significant variation from one industry to another. According to our analysis and estimate, with the continuous increase of data traffic, to satisfy the demands of the developing 3G and 4G mobile services, South Korea will experience a shortage of spectrum after 2014. Scarcity of spectrum resources means it needs to be used more efficiently. Because of legacy radio services currently operating on some of the candidate band, often with out of date technology or with an inefficient use of frequency, it is relatively difficult to recycle and adjust some of the candidate band. So, spectrum regulatory agency MSIP finalized "Mobile Gwang-gae-to Plan 2.0" for finding and supplying 1190 MHz of spectrum will be allocated for mobile communication in 4 phases by 2023 at the year-end 2013.

This study aims to find an economic impact on mobile broadband spectrum allocation with economic benefits by means of estimating service revenues and using an input-output analysis. In order to clarify the feasibility and accuracy of this study, the hypothesis involved include a few aspects as follows; Firstly, we classify scenario into the circumstances additional spectrum by increments of 700 MHz below of 3.6 GHz band. Secondly, spectrum amount of 60 MHz mobile WiMAX is excluded in the economic effect, because WiMAX and LTE are strictly divided into different market in Korea government policy. Thirdly, existing use of mobile communication spectrum is 330 MHz. Therefore, analyzing scenario allocate an additional 490 MHz to mobile broadband giving a total of 820 MHz.

Assigning additional spectrum scenario of 820 MHz into the mobile broadband service could yield an additional 40.0 trillion Won for the Korean economy in the year 2020. The newly-added benefits will be 159.6 trillion Won and creating more than 182,500 jobs over 7 years.

This paper is a pioneering study in the assessment of economy-wide effects of the mobile broadband service industries in Korea. The results provide valuable information to policy makers and decision makers by using I-O analysis. Future research needs to be undertaken to compare the marginal utility with increasing spectrum supply. Such a comparison may provide more insightful and practical results.

REFERENCES

[1] GSA, "Evolution to LTE report," GSA, June 2014.

[2] MSIP, "Wireless communication services statistics," The Ministry of Science, ICT & Future Planning (MSIP) press release, June 2014.

[3] MSIP, "Wireless data traffic statistics," The Ministry of Science, ICT & Future Planning (MSIP) press release, June 2014.

[4] MSIP, "Mobile Gwanggaeto Plan 2.0," The Ministry of Science, ICT & Future Planning (MSIP) press release, December 2013.

[5] Deloitte, "What is the impact of mobile telephony on economic growth?," November 2011, pp. 5- 6.

[6] Plum, "Valuing the use of spectrum in the EU," GSMA, June 2013, pp. 7-8.

[7] David W. Sosa and Marc V. Audenrode, "Private Sector Investment and Employment Impacts of Reassigning Spectrum to Mobile Broadband in the United States," August 2011.

[8] BCG, "The Economic Benefits of Early Harmonisation of the Digital Dividend Spectrum & the Cost of Fragmentation in Asia-Pacific," The Boston Consulting Group, May 2012.

[9] Perrin S. Meyer, Jjason W. Yungi and Jesse H. Ausubel, "A Primer on Logistic Growth and Substitution: The Mathematics of the Loglet Lab Software," Technological Forecasting and Social Change 61(3), 1999, pp.247-271.

[10] Kara Kockelman, Donna Chen, Katie Larsen and Brice Nichols Kara, "The Economics of Transportation System : A Reference for Practitioners, University of Texas at Austin, January 2013.

[11] BOK, "2009 input-output tables," The Bank of Korea, 2011.

 ISBN: 978-1-61208-358-2

Statistical Uncertainty of Market Network Structures

Petr Koldanov
Department
of Applied Mathematics and Informatics
National Research University
Higher School of Economics
Nizhny Novgorod, Russia
Email: pkoldanov@hse.ru

Panos M. Pardalos
Center for Applied Optimization
University of Florida
Gainesville, FL, USA
Email: pardalos@ise.ufl.edu

Victor Zamaraev
Laboratory of Algorithms
and Technologies for Network Analysis
National Research University
Higher School of Economics
Nizhny Novgorod, Russia
Email: vzamaraev@hse.ru

***Abstract*—A common network representation of the stock market is based on correlations of time series of return fluctuations. It is well-known that financial time series have a stochastic nature. Therefore, there is uncertainty in inferences about filtered structures in market network. Thus, market network analysis needs to be complemented by estimation of uncertainty of the obtained results. However, as far as we know there are no relevant research in the literature. In the present paper we make the first step in this direction. We propose the approach to measure statistical uncertainty of different market network structures. This approach is based on conditional risk for corresponding multiple decision statistical procedures. The proposed approach is illustrated by numerical evaluation of statistical uncertainty for popular network structures. Our experimental study validates the possibility of application of the approach for comparison of uncertainty of different network structures.**

Keywords–Statistical uncertainty; Market network model; Conditional risk; Minimum Spanning Tree; Market Graph.

I. INTRODUCTION

Network models of financial markets attract a growing attention last decades [1]–[8]. A common network representation of the stock market is based on correlations of return fluctuations. In such a representation, each stock corresponds to a vertex and a link between two vertices is estimated by sample correlation of corresponding returns. The obtained network is a complete weighted graph. In order to simplify the network and preserve the significant information, different filtering techniques are used in the literature.

One of the filtering procedures is the extraction of a minimal set of important links associated with the highest degree of similarity belonging to the Minimum Spanning Tree (MST) [1]. To construct the MST a greedy algorithm is used. A list of edges is sorted in descending order according to the weight and following the ordered list an edge is added to the MST if and only if it does not create a cycle. The MST was used to find a topological arrangement of stocks traded in a financial market, which has associated a meaningful economic taxonomy. This topology is useful in the theoretical description of financial markets and in search of economic common factors affecting specific groups of stocks. The topology and the hierarchical structure associated to it, is obtained by using information present in the time series of stock prices only.

The reduction to a minimal skeleton of links leads to loss of valuable information. To overcome this issue, Tumminello et al. [1] proposed to extent the MST by iteratively connecting the most similar nodes until the graph can be embedded on a surface of a given genus $g = k$. For example, for $g = 0$ the resulting graph is planar, which is called Planar Maximally Filtered Graph (PMFG). It was concluded by Tumminello et al. [1] that the method is very efficient in filtering relevant information about the connection structure both of the whole system and within obtained clusters.

Another filtering procedure, proposed by Boginski et al. [2], leads to the concept of Market Graph. A Market Graph (MG) is obtained from the original network by removing all edges with weights less than a specified threshold $\theta \in [-1, 1]$. Maximum cliques and maximum independent sets analysis of the Market Graph were used to obtain valuable knowledge about the structure of the stock market.

All these approaches use time series observations. It is well-known that financial time series have a stochastic nature. Therefore, there is uncertainty in inferences about filtered structures (MST, PMFG, MG) in market network. It is clear that the less numbers of observations one has the less this inferences are reliable. Thus, market network analysis needs to be complemented by estimation of uncertainty of the obtained results.

The main question is: how one can measure and compare uncertainty of different network structures, such as MST, PMFG, MG and others? To answer this question we propose to use the concept of statistical decision functions [9] and to consider statistical uncertainty. Within the framework of this approach, we introduce a measure of statistical uncertainty of market network structures. This allows to identify the most reliable network structures.

The paper is organized as follows. In Section II, we describe the approach and introduce the measure of statistical uncertainty of market network structures. In Section III, we give the results of the numerical simulations. In Section IV, we make concluding remarks.

II. MEASURE OF STATISTICAL UNCERTAINTY

Let N be a number of stocks, n be a number of days of observations. In our study financial instruments are character-

 ISBN: 978-1-61208-358-2

ized by daily returns of the stocks. Stock k return for day t is defined as

$$R_k(t) = \ln \frac{P_k(t)}{P_k(t-1)}, \quad (1)$$

where $P_k(t)$ is the price of stock k on day t. We assume that for fixed k, $R_k(t)$, $t = 1, \ldots, n$, are independent random variables with the same distribution as R_k (i.i.d.) and the random vector $R = (R_1, \ldots, R_N)$ has multivariate distribution with correlation matrix

$$||\rho_{ij}|| = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1N} \\ \cdots & \cdots & \cdots \\ \rho_{N1} & \cdots & \rho_{NN} \end{pmatrix}. \quad (2)$$

For this model we introduce the *reference network*, which is a complete weighted graph with N nodes and weight matrix $||\rho_{ij}||$. For the reference network one can consider corresponding reference structures, e.g., reference MST, reference PMFG, reference Market Graph and others.

Let $r_k(t), k = 1, \ldots, N, t = 1, \ldots, n$, be the observed values of returns. Define the *sample covariance*

$$s_{ij} = \frac{1}{n-1} \sum_{t=1}^{n} (r_i(t) - \overline{r}_i)(r_j(t) - \overline{r}_j), \quad (3)$$

and *sample correlation*

$$r_{ij} = \frac{s_{i,j}}{\sqrt{s_{i,i} s_{j,j}}} \quad (4)$$

where $\overline{r}_i = \frac{1}{n} \sum_{t=1}^{n} r_i(t)$. Using the sample correlations we introduce the (n-*period*) *sample network*, which is a complete weighted graph with N nodes and weight matrix $||r_{ij}||$. For the sample network one can consider the corresponding sample structures, e.g., sample MST, sample PMFG, sample Market Graph and others.

To handle statistical uncertainty we propose to compare the sample network with the reference network. Our comparison will be based on conditional risk connected with possible losses. The associated loss function is defined following Koldanov et al. [10] within the framework of multiple decision theory [11].

For a given structure $\mathcal{S}$, we introduce a set of hypothesis:

- h_{ij}: edge between vertices i and j is not included in the reference structure $\mathcal{S}$;
- k_{ij}: edge between vertices i and j is included in the reference structure $\mathcal{S}$.

To measure the losses, we consider two types of errors:

> **Type I error:** edge is included in the sample structure when it is absent in the reference structure;
> **Type II error:** edge is not included in the sample structure when it is present in the reference structure.

Let a_{ij} be the loss associated with the error of the first kind and b_{ij} the loss associated with the error of the second kind for the edge (i, j). According to the statistical decision theory [9] and taking into account additivity of the loss function [10], [11] we define the conditional risk for a given structure $\mathcal{S}$ as

$$\mathcal{R}(\mathcal{S}, n) = \sum_{1 \le i < j \le N} [a_{ij} P_n(d_{k_{ij}} | h_{ij}) + b_{ij} P_n(d_{h_{ij}} | k_{ij})], \quad (5)$$

where $P_n(d_{k_{ij}} | h_{ij})$ is the probability of rejecting hypothesis h_{ij} when it is true and $P_n(d_{h_{ij}} | k_{ij})$ is the probability of accepting hypothesis h_{ij} when it is false. Conditional risk is appropriate to evaluate the quality of different statstical procedures of identification of given structure. In this paper, we consider the case where $a_{ij} = 1/2M_1$ and $b_{ij} = 1/2M_2$. In this case the conditional risk is equivalent to per-family error rate (PFE) type error [12], which we call *fraction of error*:

$$\mathcal{E}(\mathcal{S}, n) = \sum_{1 \le i < j \le N} \left[\frac{1}{2M_1} P_n(d_{k_{ij}} | h_{ij}) + \frac{1}{2M_2} P_n(d_{h_{ij}} | k_{ij}) \right], \quad (6)$$

where M_1 – is a maximal possible number of type I errors and and M_2 – is a maximal possible number of type II errors.

We say that structure $\mathcal{S}_1$ *is more stable than structure* $\mathcal{S}_2$ if $\mathcal{E}(\mathcal{S}_1, n) < \mathcal{E}(\mathcal{S}_2, n)$ for any number of observations n. In other words statistical uncertainty of structure $\mathcal{S}_1$ is less than statistical uncertainty of structure $\mathcal{S}_2$ if $\mathcal{E}(\mathcal{S}_1, n_1) = \mathcal{E}(\mathcal{S}_2, n_2)$ implies $n_1 < n_2$. We define the *$\mathcal{E}$-measure of statistical uncertainty of structure $\mathcal{S}$ (of level $\mathcal{E}_0$)* as the number of observations $n_{\mathcal{E}}$ such that $\mathcal{E}(\mathcal{S}, n_{\mathcal{E}}) = \mathcal{E}_0$, where $\mathcal{E}_0$ is given value.

III. RESULTS

To illustrate our approach, we consider the network with $N = 250$ nodes and $R \sim N((0, \ldots, 0), ||\rho_{ij}^{US}||), i, j = \overline{1, N}$, where the correlation matrix $||\rho_{ij}^{US}||$ consists of pairwise correlations of daily returns of a set of 250 randomly chosen financial instruments traded in the US stock markets over a period of 365 consecutive trading days in 2010-2011. We use the matrix $||\rho_{ij}^{US}||$ as a weight matrix for our reference network. We will refer to it as the *US reference network*. Note that the only reason of this choice of the stocks is to validate our approach on a correlation matrix based on real data.

To construct the n-period sample network we simulate the sample $x_{11}, \ldots, x_{1N}, \ldots, x_{n1}, \ldots, x_{nN}$ from multivariate normal distribution $N((0, \ldots, 0), ||\rho_{ij}^{US}||), i, j = \overline{1, N}$, $N = 250$. To measure statistical uncertainty of network structure $\mathcal{S}$ we use fraction of errors $\mathcal{E}(\mathcal{S}, n)$, which we estimate in the following way:

1) In the US reference network, find reference structure $\mathcal{S}$.
2) Simulate sample $x_{11}, \ldots, x_{1N}, \ldots, x_{n1}, \ldots, x_{nN}$.
3) Calculate estimations r_{ij} of parameters ρ_{ij}^{US}.
4) In sample network (with weight matrix $||r_{ij}||$), find sample structure $\mathcal{S}$.
5) Calculate fraction of errors of type I, fraction of errors of type II and total fraction of error.
6) Repeat many times steps 1-5 and calculate $\mathcal{E}(\mathcal{S}, n)$.

In our experiments, we choose a level of statistical uncertainty $\mathcal{E}_0 = 0.1$.

 ISBN: 978-1-61208-358-2

A. Statistical uncertainty of MST

Observe that if total fraction of error $X = 0$ then reference MST and sample MST are equal; if total fraction of error $X = 1$ then reference MST and sample MST are completely different, i.e., have no common edges. The latter situation may hold for several sample MSTs under fixed reference MST. For Minimum Spanning Tree, one has $M_1 = M_2 = N - 1$, where N is a number of vertices in considered network. Note that in MST a number of errors of type I X_1 is equal to a number of errors of type II X_2. Measures of statistical uncertainty for MST can be defined from the equation:

$$\frac{1}{2(N-1)} \sum_{1 \le i < j \le N} [P_n(x_1^{ij} = 1) + P_n(x_2^{ij} = 1)] =$$

$$= \frac{1}{(N-1)} \sum_{1 \le i < j \le N} P_n(x_1^{ij} = 1) = \mathcal{E}_0, \qquad (7)$$

where $x_1^{ij} = 1$ if edge (i, j) is incorrectly included into sample structure and $x_1^{ij} = 0$ otherwise and $x_2^{ij} = 1$ if edge (i, j) is incorrectly not included into sample structure and $x_2^{ij} = 0$ otherwise. One has

$$X_1 = \sum_{1 \le i < j \le N} x_1^{ij}; \quad X_2 = \sum_{1 \le i < j \le N} x_2^{ij}; \qquad (8)$$

$$X = \frac{1}{2} \left(\frac{X_1}{M_1} + \frac{X_2}{M_2} \right). \qquad (9)$$

Results of the study of statistical uncertainty of MST are presented in Figure 1. As one can see, the condition $\mathcal{E}(\text{MST}, n) \le 0.1$ is achieved when the number of observed periods $n_{\mathcal{E}}$ is more than 10 000. Note that when $n = 1000$ sample and reference MSTs have only 70% of common edges. Moreover, by further increasing the number of observations does not lead to considerable decrease of statistical uncertainty of MST.

B. Statistical uncertainty of PMFG

Observe that $X = 0$ means that reference PMFG and sample PMFG are equal; $X = 1$ means that reference PMFG and sample PMFG are completely different, i.e., have no common edges. The latter situation may hold for several sample PMFGs under fixed reference PMFG. For Planar Maximally Filtered Graph, one has $M_1 = M_2 = 3N - 6$, where N is a number of vertices in considered network. For each edge (i, j) such that $x_1^{ij} = 1$ there is an edge (k, s) with $x_2^{ks} = 1$ and vice versa. It means that in PMFG a number of errors of type I is equal to a number of errors of type II, i.e., $X_1 = X_2$. Since M_1 and M_2 are constants, both measures of statistical uncertainty for PMFG are equivalent and can be defined from the equation:

$$\frac{1}{(3N-6)} \sum_{1 \le i < j \le N} P_n(x_1^{ij} = 1) = \mathcal{E}_0. \qquad (10)$$

Results of the study of statistical uncertainty of PMFG are presented in Figure 1. As one can see, the condition $\mathcal{E}(\text{PMFG}, n) \le 0.2$ is not achieved even when the number of observed periods $n_{\mathcal{E}}$ is equal to 10 000.

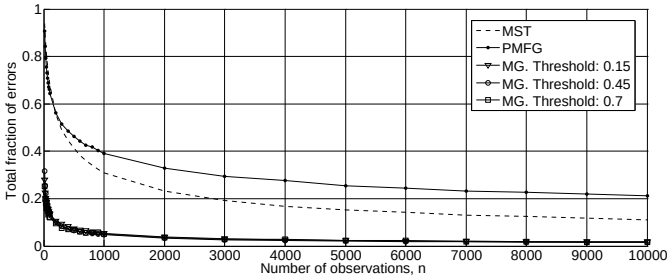


Figure 1. Total fraction of errors in PMFG, MST and MG.

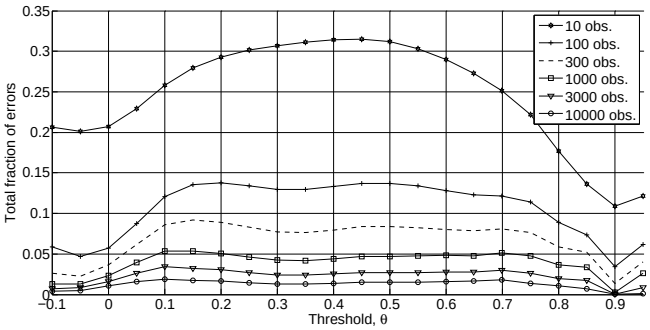


Figure 2. Total fraction of errors in Market Graphs.

C. Statistical uncertainty of MG

Observe that $X = 0$ means that reference MG and sample MG are equal; $X = 1$ means that sample MG is complement to reference MG. Let us pay attention that the latter situation for Market Graph is possible in only one case, in contrast to MST. For Market Graph one has $M_1 = \binom{N}{2} - M, M_2 = M$, where N is the number of the vertices in the considered network and M is the number of edges in the given reference Market Graph. Since M_1 and M_2 are constants, both measures of statistical uncertainty for MG are equivalent and can be defined from the equation:

$$\frac{1}{2} \sum_{1 \le i < j \le N} \left[\frac{1}{\binom{N}{2} - M} P_n(x_1^{ij} = 1) + \frac{1}{M} P_n(x_2^{ij} = 1) \right] = \mathcal{E}_0. \qquad (11)$$

Results of the study of statistical uncertainty of MG are presented in Figures 1 and 2. As one can see, the condition $\mathcal{E}(\text{MG}, n) \le 0.1$ is achieved under the number of observed periods $n_{\mathcal{E}} = 300$ for all thresholds $\theta \in [-0.1, 1]$, which is much more reasonable than the statistical uncertainty of MST.

IV. Conclusion

In the present paper, we introduced the measure of statistical uncertainty for different network structures, which is based on average fraction of errors known as per-family error rate in the theory of multiple comparison statistical procedures [12]. This measure is the particular case of conditional risk.

To illustrate our approach we consider the network where the correlation matrix consists of pairwise correlations of daily returns of a set of 250 randomly chosen financial instruments traded in the US stock markets over a period of 365 consecutive trading days in 2010-2011.

 ISBN: 978-1-61208-358-2

Our experimental study validates the possibility of application of the approach for comparison of uncertainty of different network structures. In particular, in our experiments, Market Graph is more reliable with respect to statistical uncertainty than Minimum Spanning Tree, which in turn is more reliable than Planar Maximally Filtered Graph.

REFERENCES

[1] M. Tumminello, T. Aste, T. Di Matteo, and R. Mantegna, "A tool for filtering information in complex systems," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 30, 2005, pp. 10 421–10 426.

[2] V. Boginski, S. Butenko, and P. M. Pardalos, "Statistical analysis of financial networks," Computational Statistics & Data Analysis, vol. 48, no. 2, 2005, pp. 431–443.

[3] M. A. Djauhari, "A robust filter in stock networks analysis," Physica A: Statistical Mechanics and its Applications, vol. 391, no. 20, 2012, pp. 5049–5057.

[4] M. A. Djauhari and G. S. Lee, "Minimal spanning tree problem in stock networks analysis: An efficient algorithm," Physica A: Statistical Mechanics and its Applications, vol. 392, 2013, pp. 2226–2234.

[5] S. Li, J. He, and Y. Zhuang, "A network model of the interbank market," Physica A: Statistical Mechanics and its Applications, vol. 389, no. 24, 2010, pp. 5587–5593.

[6] A. Namaki, G. Jafari, and R. Raei, "Comparing the structure of an emerging market with a mature one under global perturbation," Physica A: Statistical Mechanics and its Applications, vol. 390, no. 17, 2011, pp. 3020–3025.

[7] G. A. Bautin, V. A. Kalyagin, A. P. Koldanov, P. A. Koldanov, and P. M. Pardalos, "Simple measure of similarity for the market graph construction," Computational Management Science, vol. 10, 2013, pp. 105–124.

[8] G.-J. Wang, C. Xie, S. Chen, J.-J. Yang, and M.-Y. Yang, "Random matrix theory analysis of cross-correlations in the us stock market: Evidence from pearson correlation coefficient and detrended cross-correlation coefficient," Physica A: Statistical Mechanics and its Applications, vol. 392, 2013, pp. 3715–3730.

[9] A. Wald, Statistical decision functions. Oxford, England: Wiley, 1950.

[10] A. P. Koldanov, P. A. Koldanov, V. A. Kalyagin, and P. M. Pardalos, "Statistical procedures for the market graph construction," Computational Statistics & Data Analysis, vol. 68, 2013, pp. 17–29.

[11] E. L. Lehmann, "A theory of some multiple decision problems," The Annals of Mathematical Statistics, 1957, pp. 1–25.

[12] Y. Hochberg and A. C. Tamhane, Multiple comparison procedures. John Wiley & Sons, Inc., 1987.

 ISBN: 978-1-61208-358-2

Scalable System for Textual Analysis of Stock Market Prediction

Roy Guanyu Lin
Department of Computer Science
National Taiwan University
Taipei, Taiwan
yesimroy@gmail.com

Tzu-Chieh Tsai
Department of Computer Science
National Chengchi University
Taipei, Taiwan
ttsai@cs.nccu.edu.tw

***Abstract*—Stock Market Prediction is a problem that people deal with when they want to predict market trend. For short-term investment, news is one of the most important factors that has influence on stock price. Based on this idea, our target issue is to build a scalable stock market prediction system, which can process Chinese news articles in order to produce a prediction model. With this system, we can speed up the model training process and take into account more training source, e.g., posts from China's microblog service, Sina Weibo. Also, with the emergence of cloud computing, a scalable system can lease more resources from cloud to serve the growing work. Our solution about building this system is using mature open source project, such as Hadoop for parallel computing, Mahout for scalable machine learning, and Jieba for Chinese text segmentation. We provide a basic algorithm for stock trend prediction, build the software stack, collect the news in Taiwan during March 2009 to May 2014 and also run some experiments to evaluate scalability of this system. The result shows that in this application, Jieba Chinese text Segmentation tool can scale well with multiprocessing, namely, 80 percent faster with four parallel processes compared to sequential mode. However, Mahout does not show significant speedup in this scenario.**

Keywords-distributed system; scalability; stock market prediction

I. INTRODUCTION

Stock Market Prediction is a hot topic. There are several ways to deal with this issue. Some examples are fundamental analysis, technical analysis, hybrid analysis, and textual-based analysis. For short-term investments, news can dramatically affect stock price. One of the most famous example is the fake twitter post that Barack Obama had been injured in an explosion, causing the S&P 500 to decline 0.9%. The existing related works about textual analysis targeted the issues for chasing the prediction accuracy [1][2][3]. They use history textual information to train a prediction model, providing an algorithm to get better prediction accuracy. However, our goal is different; we focus on the scalability of Stock Market Prediction System not on the accuracy of prediction model. Because the amount of data has been exploding, we need a scalable platform to deal with large data sets to meet analysis requirement [11]. Textual analysis based stock market trend prediction needs a system with text processing function and machine learning function. However, the traditional tools are not scalable, e.g., CKIP service for Chinese text segmentation broadly used in Taiwan [22] is hard to scale. Therefore, we would like to build a scalable system using mature open source project. There are some benefits of it. First of all, this kind of system saves cost without paying any licensing fees. Second, with scalable system, we can extend the capacity of the system to deal with bigger data set for meeting user requirement, e.g. job completion within given deadline. Third, we can use cloud resource on demand to extend the capacity in pay-as-you-go manner [4].

With the emergence of cloud computing, we can bundle our scalable application into a VM image which is stored on cloud, and launch the instances from the image on demand to start the service. Also, we can adapt application capacity by configuring the amount of cloud resources leased according to workload. In this way, imagine that you have just an old laptop and an access to internet, you can still easily process a big amount of computation by using cloud resources. The amount of cloud resource you need depends on the data input. The way you pay is as you go. You do not have to buy a computer just for some temporary computations. This may save you money. However, deploying scalable system on cloud involves some issues, like how to extend the capacity from cloud, how to save cost when using the cloud resources [5][6]. Same questions appear with our platform, but these questions should be asked after the completion of system and the modeling of system performance [7].

The progress we have made is as follows. First, we create crawlers to collect the data used for stock market prediction, e.g., history news articles and history stock quotes in Taiwan. Second, we design the scalable system for stock market prediction, and build the system based on basic textual analysis based prediction algorithm. Third, we evaluate our main function components, Chinese segmentation tool and machine learning tool. The plan for system performance modeling would be our next goal after finishing the implementation of the algorithm.

The rest of this paper is organized as follows. Section 2 provides an overview description of system workflow, system software stack, and scalability issues. In Section 3, we talk about the basic algorithm of the stock market prediction. In Section 4, we show the preliminary scalability evaluation for Chinese text segmentation and classification model training. Finally, in Section 5, we illustrate future directions and make a conclusion.

 ISBN: 978-1-61208-358-2

II. OVERVIEW OF THE SYSTEM

In this section, we introduce the workflow of our system, the system software stack, and the scalability issues.

A. System workflow

We want to build a system for stock market prediction application, which we can use in Taiwan. We list some requirements for this system. First, we want to have real data to prove the feasibility of the system. To achieve this, we implemented crawlers to collect the history data, including structured data like stock prices and unstructured data like news articles. Second, because we want to make a prediction model from Chinese news articles, our system should have the ability to process Chinese Text. Similar ideas were shown in [1][2], but we deal with Chinese words, not English words. Third, we want to use machine learning technique [10] to train our prediction model, so that machines can learn how to classify a future news a good, bad, or unchanged for the company price. We will describe more details in Figure 1.

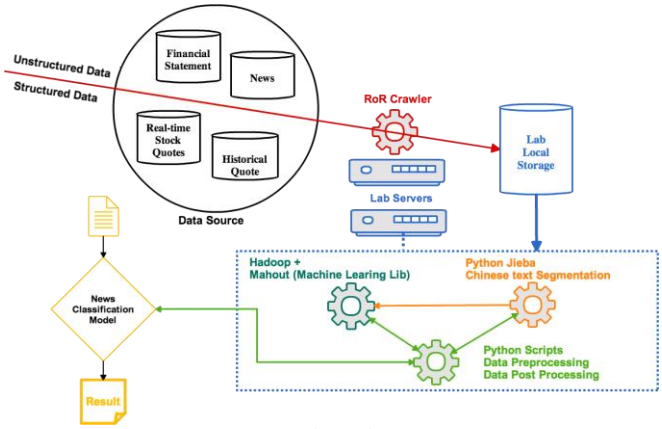


Figure 1. Workflow of the System.

First, we grab the history data from different sources, and collect these history information in order to evaluate our prediction model. We gathered news articles about Taiwan from the Internet. Because there is no good interface for collecting the news we need, we made automated crawlers using Ruby on Rails (RoR) [12] projects, such as Mechanize [13] and Watir [14]. We collected 566,114 news from 2010 to 2014, totally 2.3GB, and history stock quotes from Taiwan Stock Exchange Corporation (TSE). However, the history stock quote from TSE is daily based, which is too coarse-grain for us. We need finer-grain data if we want to design a more accurate model, so we wrote a program to record per-minute based stock quotes. Based on these two history information, news articles and stock quotes, we can do our prediction model training. The model training process will be described in section three. In this process, we need some tools dealing with Chinese Text Segmentation, classification model training, and also scripts for data pre-processing and post-processing. After we get a news classification model, once a news appear, the system will be triggered, and output whether the news make the company price go up, down, or stay unchanged. The next part will explain the software stack we plan to build. Then, some scalability issues will be discussed.

B. System Software Stack Design and Implementation

Our system architecture is presented in Figure 2. The orange part are the local resources, which can be physical servers or virtualized servers. The purpose of this design is to provide a better utilization of the physical servers. The blue part stands for cloud resources, and we take Amazon for example as our service provider. Amazon Web Services (AWS) [15] provides a lot of web services. Amazon EC2 [16] is one of them, which belongs to IaaS service model. Cloud computing has three service models, e.g., infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) [9]. EC2 provides lots of different specifications of virtual machine to customers. Amazon S3 [17] provides key-value storage service, which can be easily integrated with other products within Amazon. Amazon Elastic MapReduce is the service which offer better abstraction, omitting the steps of building a MapReduce runtime environment. The red part, is a famous scalable Hadoop ecosystem.

We choose Hadoop ecosystem to meet machine learning tool requirement. Mahout is the machine learning library, which is also an Apache project [18], resides on the Hadoop MapReduce stack [19]. This project has three main functions, namely, recommendation, classification, and clustering. In this paper, we use the function of classification. Last, for the green part, it means those scripts for data pre-processing and post-processing, and RoR crawlers.

In the current progress, the crawler implementation have already finished, and we collect 566,114 articles of 2.3GB in size. The software stack has been setup. However, the implementation of the basic stock market prediction is still under development. Which the algorithm will be explained later.

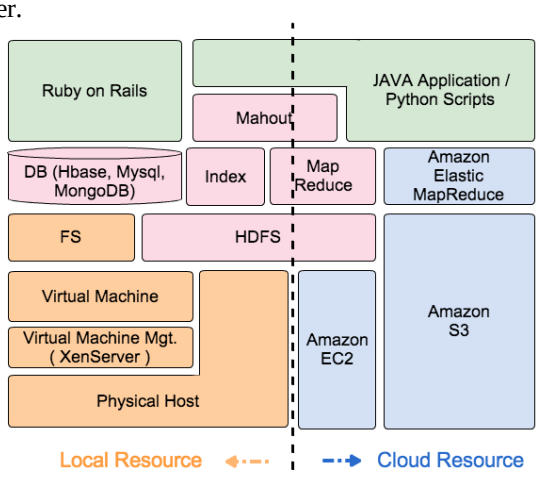


Figure 2. System Software Stack.

 ISBN: 978-1-61208-358-2

C. Scalability Issues

Scalability is the ability of a system to handle a growing amount of work. We parallel our computation tasks to get system scalability. Chinese segmentation is one of the main function components, and we discuss scalability of this part as our first step. In our system, data is stored in Hadoop Distributed File System (HDFS). HDFS provides file replication, which can maintain several copies on different machines. Also, due to the nature of data independence for each news articles, the computation can be parallelized. In short, with data replication and data independence, we can parallel Chinese segmentation jobs on those nodes with data copies.

Jieba project [20] originally provides a module to parallel Chinese segmentation process. It cuts a file into several lines, and distributes the lines to several workers in order to increase the throughput. However, when we use this module on our 566,114 Chinese news articles, sized 2.3G, the response time increases from 68 minutes to 99 minutes, shown in Figure 3. Figure 3 compares the completion time of a Chinese segmentation job between sequential mode and parallel mode with 2, 4, and 24 workers. The reason is that each news article is short, the overhead of separating lines to workers is higher than the benefit of parallel computation on multiple workers. Therefore, we change the way from line parallelism to file parallelism, distributing news article files into many workers (processes), as shown in Figure 4.

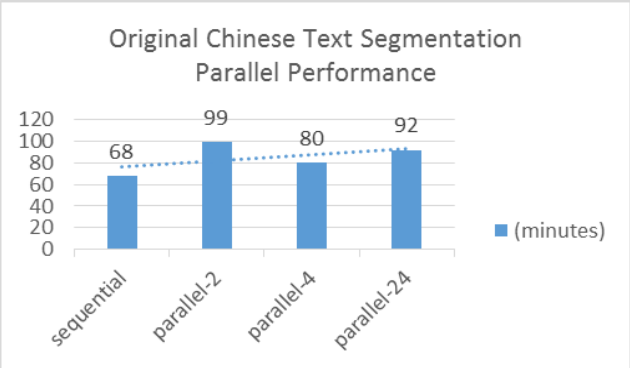


Figure 3. Performance Evaluation of Original Multi-Processes Scaling Up.

In section four, we evaluate the performance of scaling up (vertical scaling) and scaling out (horizontal scaling) of file parallelism version of Chinese text segmentation process using Jieba tool, and describe a problem we met for classification model training using Mahout.

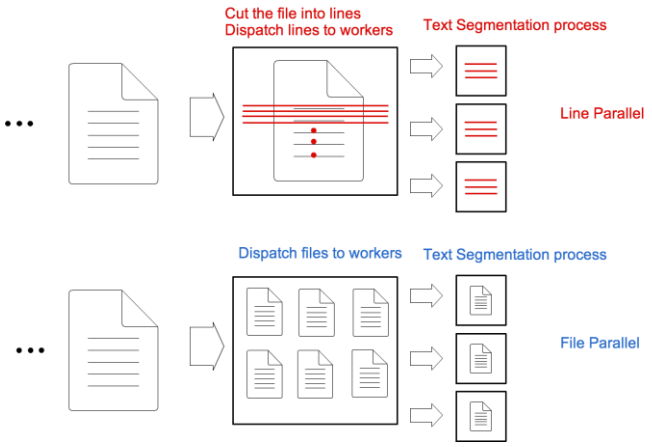


Figure 4. Original Jieba Line Parallelism and our Inter-File Parallelism.

III. BASIC STOCK MARKET PREDICTION ALGORITHM

In this section, we introduce our basic market prediction algorithm. Now, we use Chinese text articles to predict whether the stock price goes up, down, or stay unchanged of our target companies. In the future, we will take social media information, e.g., microblog posts, into consideration. The implementation of the algorithm is still under development. By building this algorithm we would like to prove the feasibility of our system for stock trend prediction.

The overall process of the system is as follows. The process consists of two parts. The first part is training the news classification model, and the second part is using the classification model to classify new text articles for predicting the stock trend.

For the second part, assume that we already have a classification model. When a news appears, news sensor detects the events and triggers news classification for each company to see whether this article makes the company stock price up/down, or just stay unchanged. How we get the model of news classification is described below.

The process of training classification model described in Figure 5. At the beginning, with a target company, the script automatically gets history stock quotes from database, filters dates by variation. The variation now is set to 25% variation for an hour. If the stock price goes up over 25% between time points, then we label the time interval “up”, vice versa. If it is between -25% and +25%, we label the time interval “unchanged.” In this way, we can get time intervals labeled “up”, “down”, or “unchanged”. Then the script searches the news articles related to the company with time and label input, and then tags the articles with label “up”, “down”, or “unchanged”. After labeling, the script triggers Chinese text segmentation. Also, we provide our customized Chinese dictionary to make the segmentation more accurate and do noise filtering. At last, we input the article set with featured words and labels into Mahout Naïve Bayes classification training process [21]. Before training, the script splits

 ISBN: 978-1-61208-358-2

dataset into training set and test set. After we get a model, the script evaluates the model by test set and reports the prediction accuracy of classification model.

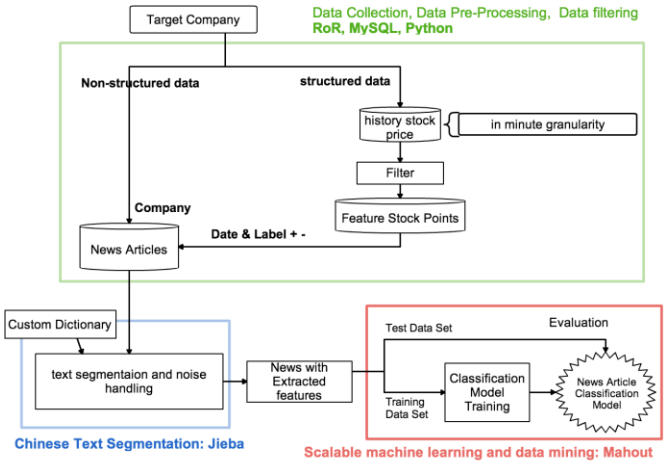


Figure 5. Model Training Process for News Classification.

IV. PRELIMINARY EXPERIMENTS

We did some experiments for evaluating the scalability of file-parallelism version of Jieba and Classification tool Mahout. The experiment environment is using Ubuntu 12.04 OS, running on one 24 cores 16G RAM server, and three 8 cores 8G RAM server.

We use Jieba Python version to process 2.3G Chinese news, totaled 566,114 articles. We use process pool to create many parallel processes to segment the news articles. Figure 6 shows the result of scale-up performance improvement on the 24 cores server. Compared to sequential version, four processes parallelism is 80% faster. In addition to scale-up (vertical scaling) experiments, we still made scale-out (horizontal scaling) experiments. Figure 7 shows the scale-out experiment of processing 2.3G news articles on the 8 cores 8GB server in sequential mode. We split data into two copies and three copies using scripts to three servers, and test the performance. As our expectations, the performance improvement is almost linear; for the constant module, loading time is small compared to workload computation. In the future, we think about integrating job parallelism with HDFS. HDFS default stores three copies for every chunk; so, we do not need to write one more script to deal with the data split action.

We did several experiments to test the scalability of Mahout. The experiment is conducted on one node, two nodes, and three nodes runtime environment. However, we could not get a significant performance improvement. To find out the reason, we decompose the auto script into small steps for Mahout Naïve Bayes Classification and record its latency. We found that the first step, seqdirectory, the command of which makes the files in HDFS sequential, always produces just one map task in the job. It means we cannot parallel the computation in this step. Usually, the number of map tasks is related to the number of chunks in HDFS. Our data size is more than 640MB. If it combines all the small files into a big one, it should at least have 10 chunks with default chunk size 64MB. We have not found a solutions yet. We tried to configure Hadoop several times, but failed. Now, we are still tracing Mahout source code for solving the problem. The latency of every steps in Naïve Bayes classification is depicted in Figure 8.

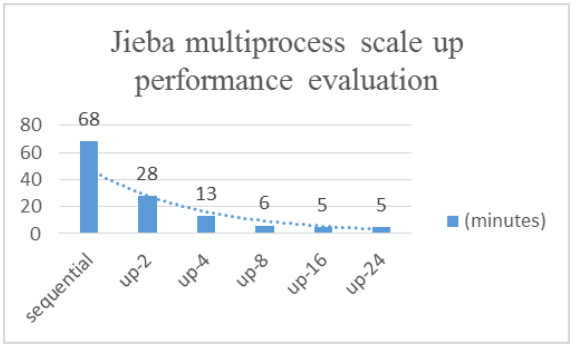


Figure 6. Performance Evaluation of File Parallelism Scaling Up.

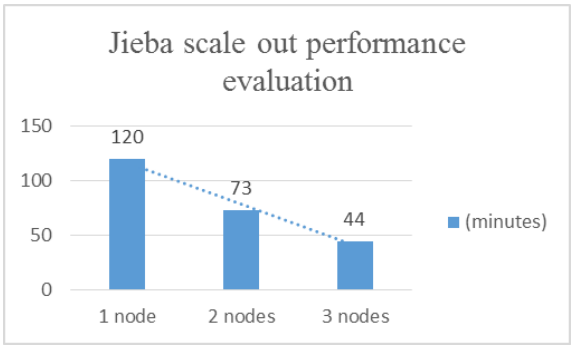


Figure 7. Performance Evaluation of File Parallelism Scaling Out.

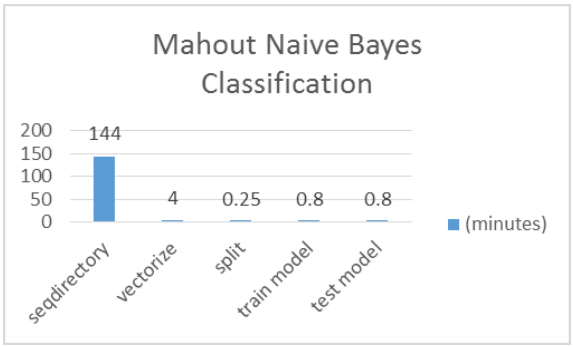


Figure 8. Performance Evaluation of Mahout Naive Bayes Classification.

V. FUTURE DIRECTION AND CONCLUSION

In this paper, after the preliminary result, we proved that Jieba can scale well with our file parallelism version, i.e., scaling-up with four cores gets 80% faster compared to one core environment and scaling-out makes linear improvement. The main factors of scalability are the nature of data independence and data replication. Text segmentation

 ISBN: 978-1-61208-358-2

process can be executed in a parallel way. However, the performance improvement of Mahout is limited by the first step, i.e., file sequential process. Because the bottleneck has huge impact on overall response time of classification model training process, we have to deal with it in the future. We are still solving this problem by tracing Mahout source code. It is worth mentioning that Mahout starts to move its focus on Spark [8], a new popular large scale data processing project stating faster processing speed because of in-memory computation. We will use Spark to solve the scalability issue of machine learning function in another way.

In addition to Mahout and Jieba, we will also evaluate another components in our system to prove scalability as soon as we finish building our system. Also, we will use queueing theory to build system performance modeling. With performance model, we can adapt system resource to make performance meet user requirements. Also, we will consider the issues about offloading to cloud. For example, when will we need extra resources, how to offload computations to cloud, and how to use cloud resources in a cost-aware way.

Although the work is not finished yet, we believe this is a good issue worth discussing. The era of big data is coming, a scalable system for this kind of application is needed. Because we may develop new prediction algorithm based on bigger data source, e.g., social media information, with the sharing of the experience, we believe it is helpful to give readers a hint to build a scalable system for textual analysis based stock market trend prediction.

REFERENCES

[1] Fung, Gabriel Pui Cheong, Jeffrey Xu Yu, and Wai Lam. "Stock prediction: Integrating text mining approach using real-time news." Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on. IEEE, 2003.

[2] Schumaker Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) 27.2 (2009): 12.

[3] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. "Mining of concurrent text and time series." KDD-2000 Workshop on Text Mining. 2000.

[4] Armbrust, Michael, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee et al. "A view of cloud computing." Communications of the ACM 53.4 (2010): 50-58.

[5] Sharma, U., Shenoy, P., Sahu, S., & Shaikh, A. (2011, June). "A cost-aware elasticity provisioning system for the cloud." Distributed Computing Systems (ICDCS), 2011 31st International Conference on. IEEE, 2011.

[6] Guo, T., Sharma, U., Wood, T., Sahu, S., & Shenoy, P. J. "Seagull: intelligent cloud bursting for enterprise applications." Proceedings of the Usenix Annual Technical Conference (short paper). 2012.

[7] Ganapathi, A., Chen, Y., Fox, A., Katz, R., & Patterson, D. "Statistics-driven workload modeling for the cloud."Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on. IEEE, 2010.

[8] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. "Spark: cluster computing with working sets."Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. 2010.

[9] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011).

[10] Naïve Bayes Classification. (2014)Retrieved July 3 , 2014, from http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

[11] Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).

[12] Ruby on Rails Web Framework. (2014). Retrieved July 2, 2014, from http://rubyonrails.org/

[13] Mechanize Ruby Gem. (2014). Retrieved July 2, 2014, from https://rubygems.org/gems/mechanize

[14] Watir Ruby Gem. (2014). Retrieved July 2, 2014, from https://rubygems.org/gems/watir

[15] Amazon Web Service. (2014). Retrieved July 3, 2014, from http://aws.amazon.com/

[16] Amazon Web Service Elastic Compute Cloud (EC2). (2014). Retrieved July 3, 2014, from http://aws.amazon.com/ec2/

[17] Amazon Web Service S3. (2014). Retrieved July 3, 2014, from http://aws.amazon.com/s3/

[18] Apache Mahout Project. (2014). Retrieved July 4, 2014, from https://mahout.apache.org/

[19] Apache Hadoop Project. (2014). Retrieved July 4, 2014, from http://hadoop.apache.org/

[20] Jieba Project for Chinese Text Segmentation. (2014). Retrieved July 4, 2014, from https://github.com/fxsjy/jieba

[21] Mahout Naïve Bayes. (2014). Retrieved July 5, 2014, from https://mahout.apache.org/users/classification/bayesian.html

[22] CKIP Chinese Text Segmentation Tool. (2014). Retrieved July 3, 2014, from http://ckipsvr.iis.sinica.edu.tw/

 ISBN: 978-1-61208-358-2

Evolutionary Clustering Analysis of Multiple Edge Set Networks used for Modeling Ivory Coast Mobile Phone Data and Sensemaking

Daniel B. Rajchwald, Thomas J. Klemas
Network Science Research Centre
Swansea University, Swansea, Wales
Emails: rajchwa1@gmail.com, tklemas@alum.mit.edu

***Abstract*—Static and evolutionary clustering approaches exist that enable dynamically adaptive cluster analysis of large networks. These techniques are typically based on any of the traditional techniques, such as k-means, spectral, Kerninghan-Lin, and other partitioning or clustering algorithms. In this paper, we utilize spectral clustering and k-means as the fundamental clustering mechanisms but combine adaptive and evolutionary clustering to capture problem dynamics. We apply our approach to analyze a complex, dynamic multiple edge set network that was used to model call data from the Ivory Coast compiled from France Telecom/Orange anonymized call records over a 5 month period. Our methods are used to identify important but non-evident structural groupings, resolve community clusters, develop insights based on the evolving structure and associated history, and to make sense of the raw data, the ultimate objective for Sensemaking technologies.**

Keywords–Sensemaking; adaptive clustering; spectral clustering; network theory; silhouette; k-means.

I. INTRODUCTION

Large data sets frequently contain patterns that are difficult to discern through observation alone. Data points in large data sets are often grouped in one or more dimensions based on similarities in data point values along those dimensions. However, many tools exist to group data based on proximity measures, such as Euclidean distance (where applicable), silhouette values, Saltines cosines, Pearson coefficients, or other measures of equivalence [1] that help evaluate similarity or dissimilarity between data points. Data clustering based on traditional algorithms, such as k-means, Spectral clustering, and Kerninghan-Lin, or hybrid combinations of these methods, can be a valuable tool to gain insight into many different types of data sets [1]. Once clusters are determined, new measurements can be classified more quickly based on proximity measures and parameters of the known clusters. However, over time, data groupings, clusters of data points, or even fundamental underlying network structure can evolve resulting in drift of parameters of the associated proximity measures. There has been significant study of cluster drift and related concepts of incremental and constrained clustering [2][3][4]. Evolutionary clustering techniques have been developed to capture cluster drift into clustering algorithms yet resist unduly perturbing clustering based on noise within the data by incorporating notions of expected smoothness in cluster parameters [2]. We adapt these concepts to accomplish evolutionary clustering analysis of a multiple edge set network used to model the Ivory Coast France Telecom/Orange call records.

The rest of this paper is organized as follows. Section II describes the technical details of adapting the evolutionary clustering algorithms for analysis of a multiple edge set network. Section III describes details of the data set and applying the algorithms to this data set. Section IV presents results of the multiple edge set network evolutionary clustering analysis. Section V presents our conclusions. The acknowledgement and references close the article.

II. TECHNICAL DETAILS

Now, we introduce our notation and review the basics of clustering. We model the social network derived from the call data by a graph G comprised of vertices V and edges E_1 and E_2 that represent subprefectures and the 2 types of connections between them, respectively.

$$G = (V, E_1, E_2) \tag{1}$$

The edges that connect vertex pairs represent calls, E_1, or travel, E_2, between those 2 paired subprefecture vertices. Even though the call and travel records identify the originating and terminating nodes in an edge, we construct an undirected graph model for simplicity. These cell towers are geographically distributed throughout the various subprefectures of the Ivory Coast, so there is an additional layer of mapping between the cell tower nodes and subprefecture nodes to which we applying the clustering analysis. A community S_i is comprised of a cluster of nodes, disjoint to every other community, because no vertex exists in more than one community.

$$V = \bigcup S_i, \forall_{i,j,i\neq j} S_i \cap S_j = \emptyset \tag{2}$$

To cluster the subprefectures into communities, we can assign each subprefecture, a, a feature vector, f_a and directly cluster the feature vectors into k clusters using the k-means algorithm. A more robust approach, [5] computes spectral decomposition

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{3}$$

of the $N \times N$ similarity matrix, $\boldsymbol{W}$ that is derived from the feature vectors, f_a,

$$W_{ij} = e^{\frac{-|(f_i - f_j)|^2}{2\sigma^2}} \tag{4}$$

and then clusters the row space of the eigenvectors, $\boldsymbol{U}$, corresponding to the largest k eigenvalues, by applying the k-means algorithm to the k-element rows of $[U_1 ... U_k]$ to compute k clusters. In [5], the parameter σ^2 determines the decay of

 ISBN: 978-1-61208-358-2

the affinity matrix values with distance in the feature space. In our implementation, we calculate σ^2 as the sample sum of the variances of each feature.

For evolutionary clustering, the silhouette metric is then used to determine the strength of community structure in an induced clustering, where the silhouette value [6] of one node or subprefecture, i, is defined as

$$silhouette\,(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{5}$$

where $a(i)$ is the average dissimilarity of i with the other subprefectures in its cluster and $b(i)$ is the minimum of the dissimilarities of i with all clusters that do not include i. Dissimilarity between 2 subprefectures is measured by the distance between their feature vectors. For instance, the dissimilarity between 2 feature vectors can simply be the Euclidean distance. The silhouette value for the entire clustering of nodes into k communities is simply the mean of the node silhouette values

$$silhouette(k) = mean_i(silhouette(i)) \tag{6}$$

Note that silhouette values range from -1 to 1, where 1 represents a strong community structure, -1 represents weak community structure, and 0 represents that the induced clustering is on the border with another viable clustering.

Then, we use a modified version of spectral clustering [2] to add temporal smoothness to clusterings. Before the modified version is discussed, we define some basic quantities. Given two subsets V_1 and V_2 of node set V, the association between the two subsets is $assoc(V_1, V_2) = \sum_{i \in V_1, j \in V_2} W(i,j)$ and the k-way average association between k clusters is $AA = \sum_{l=1}^{k} \frac{assoc(V_l, V_l)}{|V_l|}$. The modified spectral clustering [2] minimizes negated average association cost between two clusterings in adjacent time steps defined by

$$Cost_{NA} = \alpha NA_t|_{Z_t} + \beta NA_{t-1}|_{Z_t} \tag{7}$$

$$NA = Tr(W) - \sum_{l=1}^{k} \frac{assoc(V_l, V_l)}{|V_l|} = Tr(W) - Tr(Z^T W Z) \tag{8}$$

to obtain a clustering of the nodes at time t that is consistent with the network at time $t-1$. α and β, $\alpha + \beta = 1$, define the snapshot and temporal weights, respectively. Z_t is the $n \times k$ matrix that defines the partitioning at time t where $Z(i,j) = 1$ if and only if node i belongs to cluster j. Substituting equation 8 into equation 7 yields

$$Cost_{NA} = Tr(\alpha W_t + \beta W_{t-1}) - Tr(Z_t^T(\alpha W_t + \beta W_{t-1})Z_t) \tag{9}$$

Minimizing $Cost_{NA}$ is equivalent to maximizing $Tr(Z_t^T(\alpha W_t + \beta W_{t-1})Z_t)$ and optimizing Z_t turns out to be equivalent to applying spectral clustering to $W = \alpha W_t + \beta W_{t-1}$ [2], where W is the similarity matrix used in equation 5. Thus, by applying k-means to the rows of the matrix containing k eigenvectors corresponding to the top k eigenvalues of $W = \alpha W_t + \beta W_{t-1}$ yields the clustering at time t that maximizes both the snapshot and temporal quality.

III. Application to Data Set

A. Description of Data Set

This section of the paper is based on Blondel and Esch [7]. The data was organized into multiple sets. This research focused on Data Sets 1 and 2 in the Data For Development (D4D) collection. Data Set 1 consisted of antenna to antenna call records that include number of calls and duration of calls between any pairs of antennas, accumulated for each hour. Data Set 2 consists of records that identify cell phone tower indices for 500,000 randomly sampled callers but provided for only a 2 week duration. Data Set 3 consists of records that identify subprefecture indices for 50,000 randomly sampled callers for the entire 5 month duration of the D4D data. We decided to use Data Set 2 instead of Data Set 3 to model the traveler activity since the tower communication was recorded on a tower to tower basis. The data set also includes additional files that provide geographical location of antennas and subprefecture geographical center locations, enabling a mapping between antennas and the nearest subprefectures centers and thereby a graphical geographical depiction of result data.

B. Applying the Algorithms

We cluster the 255 subprefectures using temporal information with antenna call and/or cell phone user data. For example, if the feature vector was constructed entirely by antenna calls and time, then the feature vector for a would be defined as follows

$$f_a(t,b) = nCalls(t,a,b) \tag{10}$$

where $nCalls(t,a,b)$ represents the length of total cell phone tower communication between subprefectures a and b over a time period t. We then cluster all the feature vectors using spectral clustering as implemented by Ng and Jordan [5] (except we do not set the diagonals of the similarity matrix to 0) and the standard k-means algorithm. Given that k-means clustering does not account for noise and correlations in data, we quantify spectral clustering's effectiveness in clustering noisy and correlated data by comparing the performance of the two approaches. We implement the algorithms for cluster numbers, k, from 2 to 12 and compute the silhouette value of each clustering. The upper cluster number 12 was empirically determined from silhouette values that yield low numbers beyond 10 clusters. Once an optimum clustering is obtained, one can make inferences about relationships between the clustering features and established Ivory Coast information such as geographical, cultural, and political facts.

To compare the similarity of two clusterings, C_1 and C_2, over the same network, we first define the similarity score of a node, i, to be

$$SimilarityScore(i) = \frac{|C_1(i) \bigcap C_2(i)|}{|C_1(i) \bigcup C_2(i)|} \tag{11}$$

where $C_1(i)$ and $C_2(i)$ denote i's community in clusterings C_1 and C_2, respectively. Taking the mean of the similarity score over all nodes in the network yields the similarity score of the two clusterings, $SimilarityScore(C_1, C_2)$.

IV. Results

A. Clustering on Antenna Communication Edge Set Network

Cluster analysis was accomplished using feature vectors representing antenna communication between subprefectures

 ISBN: 978-1-61208-358-2

over 1 week from February 3, 2012 to February 9, 2012. We used two different sets of feature vectors. We defined the first set of feature vectors to be the cumulative activites of subprefectures over the 1 week period. For this section, we define activity to be antenna communication. Thus, a feature vector of a single subprefecture, a, would be a 255 dimensional vector, g_a, where $g_a(b)$ is the total length of calls beween subprefectures a and b accumulated over the entire week. We defined the second set of feature vectors to be the dynamic or evolutionary activity of the subprefectures over the 1 week period with a time interval of 24 hours. Thus, the feature vector for subprefecture a would be f_a where $f_a(b, n)$ is the total length of calls between subprefectures a and b on day n. Although the Ivory Coast regions have different populations, there is no population data on individual subprefectures so we were not able to normalize the feature vectors by population. Clustering was implemented using spectral clustering and k-means on the two sets of feature vectors. Figure 1 shows the result of spectral clustering on the dynamic subprefecture activities using 3 clusters.

We then applied adaptive spectral clustering to each of the 7 days from February 3, 2012 to February 9, 2012 as adopted from [2], for cluster sizes k = 2 to 12. A feature vector for a subprefecture, a, for a single day, n, would be defined as $h_a(b) = f_a(b, n)$ for f_a defined above. We used a snapshot cost, α, of 0.8 and a temporal cost, β, of 0.2. We computed the similarity score for each day with respect to both the cumulative and dynamic antenna communication activity clusters (through spectral clustering) over the February 3 to 9 interval. The similarity score of a day, n, was computed by averaging the similarity score of n's 11 clusters from k = 2 to 12 with either the evolutionary or cumulative clustering for February 3 to 12. The evolutionary and cumulative clusterings for the week serve as a common average to compare to each day. The results are plotted in Figure 2. Note that the day similarity scores are all relatively high and curves show oscillatory patterns.

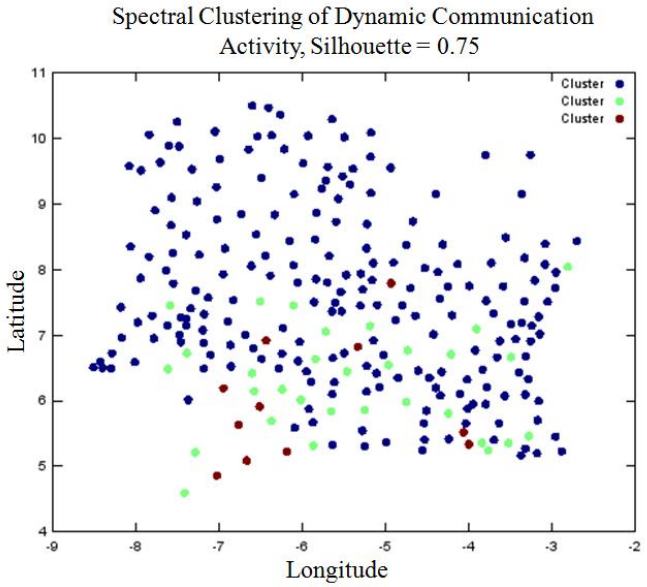


Figure 1. Example of Spectral Clustering using Subprefecture Communication Data

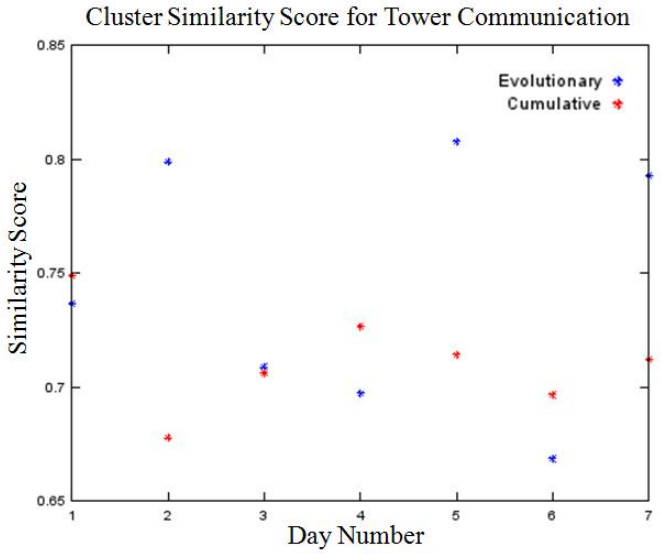


Figure 2. Similarity between Single Day Clusterings and Average Clusterings for the Week (Networks formed from Communication Data)

B. Clustering on Travel Edge Set Network

Cluster analysis was also achieved using the edge set corresponding to travel between subprefectures using spectral clustering and k-means. The second data set provides the location and times of cell phone users throughout the Ivory Coast. By geolocating the cell phone users by subprefectures, one can track when and where they travel between subprefectures. The cumulative and dynamic feature vectors were also formed from traveler data (D4D Data Set 2) between February 3, 2012 and February 9, 2012 (the dynamic travel vectors also with a time increment of 24 hours). The result of applying k-means when k = 3 to the dynamic traveler data can be seen in Figure 3. Despite the nice appearance of the 3 tight clusters, the clustering had a low silhouette score of -0.26.

Over all 12 spectral clusterings, the one with k = 2 clusters yielded the highest silhouette value in cases of antenna communication activity, as described in the last section, and traveler activity. Both k = 2 clusterings isolated the red subprefecture are seen in Figure 4. This subprefecture corresponds to Abidjan's location, the largest city and economic center of the Ivory Coast (355 of the 1031 cell phone towers were mapped to this subprefecture). Abidjan's prominant role would explain why its cell phone tower communication and traveler data are very different from other subprefectures. This does not imply that the other 254 subprefecture are similar, just that none of them are similar to Abidjan. In all clustering algorithms we used, we ran k-means numerous times with different initial random centroid placements to ensure Abidjan is a true singlet cluster. The silhouette scores of the remaining clusterings will be discussed in the next section. Figure 5 shows the similarity score for each day of the week using traveler activity in the feature vectors. The similarity scores are all higher than corresponding antenna communication similarity scores shown in Figure 2, indicating that travel does not seem to change as much as antenna communication from day to day. Both evolutionary and cumulative curves follow the same oscillatory pattern, unlike in Figure 2, implying that cumulative and dynamic traveler behavior are more consistent than the

 ISBN: 978-1-61208-358-2

corresponding communication behavior.

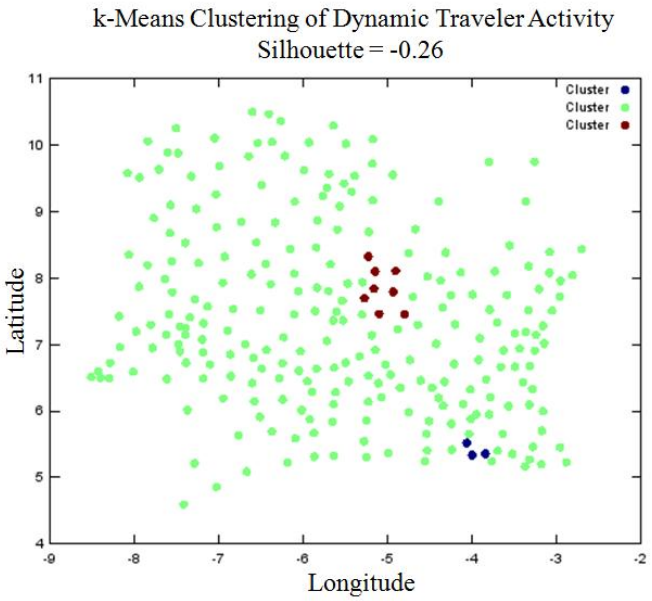


Figure 3. Example of k-means using Subprefecture Travel Data

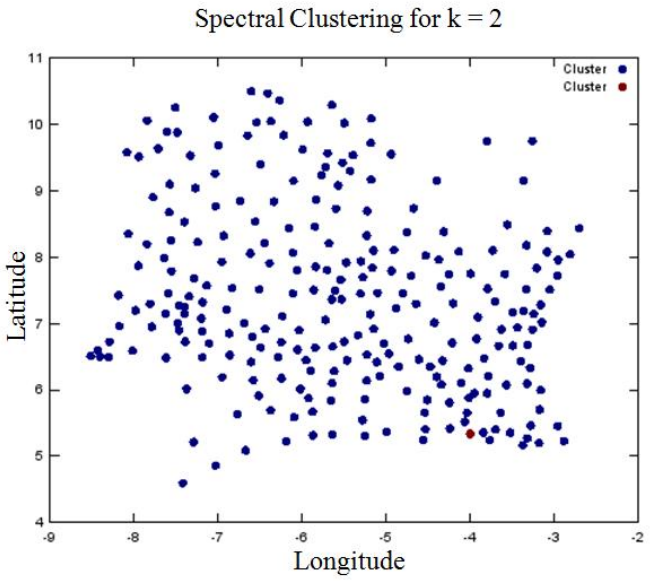


Figure 4. Spectral Clustering when k = 2 in cases of Subprefecture Communication and Travel Data

C. Multiple Edge Set Clustering

We concatenated the antenna communication and traveler feature vectors in the previous two sections to see the effect of our clustering methods on a network with more than one edge type. In Figure 6, we see the similarity scores computed for February 3 to February 9 using combined tower communication and traveler features. Both evolutionary and cumulative curves show less noisy patterns than in Figures 5 and 2 and we see a clear dip and minimum for both curves at Day 4. In Figures 7 and 8, we plot the silhouette values for each number of clusters for each combination of clustering

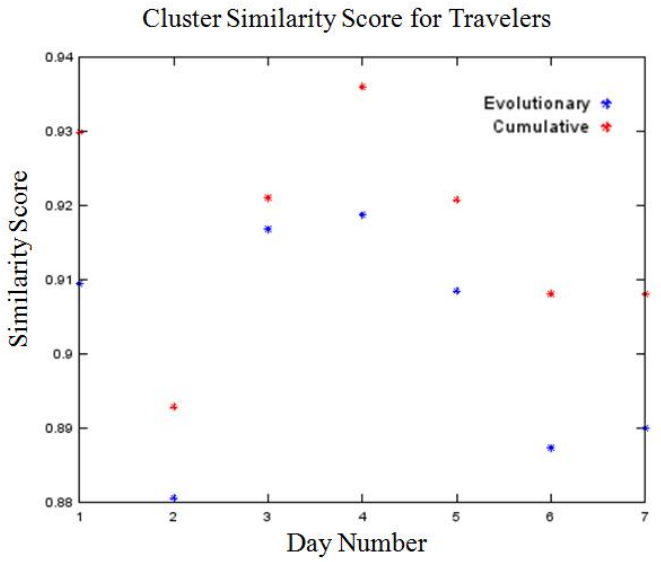


Figure 5. Similarity between Single Day Clusterings and Average Clusterings for the Week (Networks formed from Traveler Data)

algorithm and feature type. In Figure 7, we cluster cumulative activity and in Figure 8, we cluster the dynamic activity, both from February 3 to 9. Both plots are very similar, showing there is little difference between the strength of community structure between cumulative and evolutionary activity. There is a dramatic increase in silhouette values, the green curves, for the travel data from k-means to spectral clustering in Figures 7 and 8. The same is not true for the communication features and the combined communication and travel features (the red and blue curves). For both the red and blue points, spectral clustering silhouette values are very close to k-means silhouette values though there is marginal improvement when the number of clusters is more than 6. The improvement of spectral clustering over k-means depends substantially on the geometry of the feature values [5]; so, it is likely the case that the geometry of the antenna communication feature vectors is more conducive to k-means than the traveler feature vectors.

V. Conclusion

We applied clustering techniques to antenna communication and traveler data from February 3 to 9, 2012 between 255 Ivory Coast subprefectures. The optimum clustering for all feature and clustering algorithm combinations occurs when the number of clusters is 2 due to the unique central position of Abidjan (see Figure 4). While the cluster similarities scores were relatively high throughout the week in all cases, there was a smoother pattern seen in the case when communication and traveler features are combined though more work needs to be done to verify the cause of this. The consistency of the community structure over time can also be seen through the proximity between the strengths of the evolutionary and cumulative community structures (see Figures 7 and 8). Spectral clustering dramatically improved the community structure over k-mean clustering in the traveler feature space. However, the community structure over the combined traveler/communication feature space is only marginally better on average than that over the communication feature space. By adapting dynamic clustering techniques for networks with multiple edge sets,

 ISBN: 978-1-61208-358-2

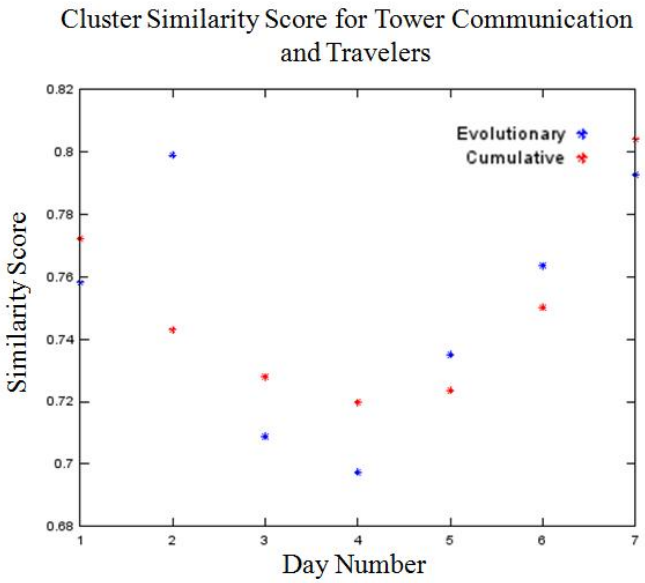


Figure 6. Similarity between Single Day Clusterings and Average Clusterings for 2/3 to 2/9 (Networks formed from Communication and Traveler Data)

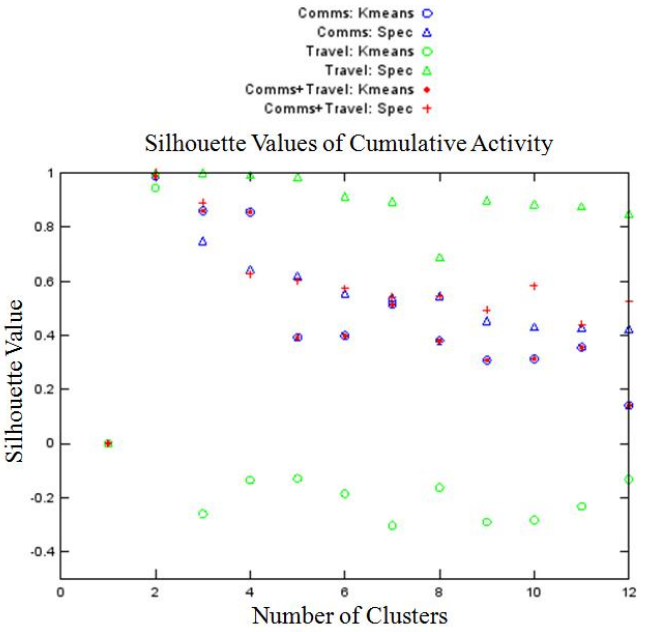


Figure 7. Silhouette Values of Cumulative Activity (2/3 to 2/9) Clusterings over Different Data Features and Algorithms

we were able to make sense of key spatial and temporal network attributes and propose new questions about clustering in heterogeneous networks.

ACKNOWLEDGMENT

The authors would like to thank the Sensemaking/PACOM Fellowship and Swansea University's Network Science Research Center for providing inspiration that led to this research and the opportunity to analyze this data set. Finally, we would like to thank Dr. Steve Chan, Jan-Kees Buenen, and Stef van den Elzenhave for advice regarding this paper.

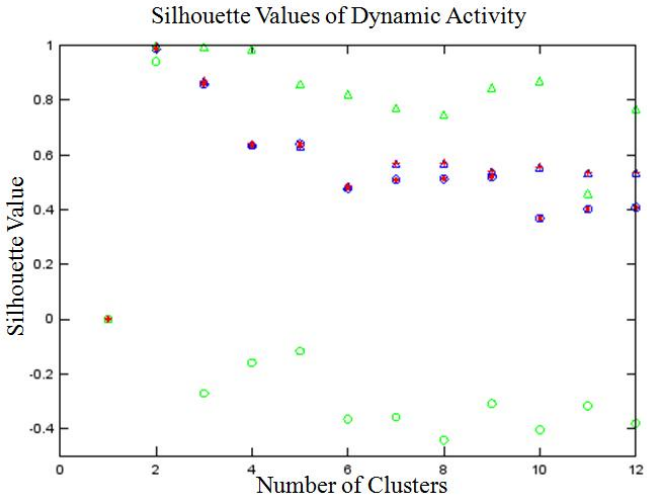


Figure 8. Silhouette Values of Dynamic Activity (2/3 to 2/9) Clusterings over Different Data Features and Algorithms

REFERENCES

[1] M. Newman, Networks, An Introduction. Oxford: Oxford University Press, 2010.

[2] Y. Chi, X. Song, D. Zhou, K. Hino, and B. Tseng, "Evolutionary spectral clustering by incorporating temporal smoothness," in KDD Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 2007, pp. 1–5.

[3] P. Grindrod and D. Higham, "Evolving graphs: Dynamical models, inverse problems, and propagation," in Proceedings of the Royal Society A, 2009, pp. 753–770.

[4] H. Jo, R. Pan, and K. Kaski, "Emergence of bursts and communities in evolving weighted networks," PLOS ONE, 2011, pp. 1–3.

[5] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in Neural Information Processing Systems (NIPS), vol. 14, 2002, pp. 1–6.

[6] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Computational and Applied Mathematics, vol. 20, 1987, pp. 53–65.

[7] V. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, and E. Huens, "Data for development: The d4d challenge on mobile phone data," arXiv:1210.0137v1 [cs.CY], Sep 2012, pp. 5–9.

 ISBN: 978-1-61208-358-2

Property Preservation in Reduction of Data Volume for Mining: A Neighborhood System Approach

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

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

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

Matthew Antonelli and Bryan Dahlqvist
Department of Computer Science
Armstrong State University
Savannah, GA, USA
Mattr.Antonelli@gmail.com
n_bryan28@hotmail.com

***Abstract*— The sheer volume of the very large datasets is the major obstacle in mining of the data because the size of the dataset is above the handling abilities of the traditional methodologies. A considerable vertical reduction over and beyond the reduction prescribed by pre-mining processes is needed to overcome the problem. However, the reduced version of the dataset ought to preserve the intrinsic properties of the original dataset in reference to a specific mining goal (a robust reduction); otherwise, it is a useless reduction. This research effort introduces and investigates the neighborhood system as a robust data volume reduction methodology in reference to the mining goal of "prediction". Two well-known prediction algorithms of ID3 and Rough Sets are employed to determine the perseveration of intrinsic properties in the reduced datasets. The results obtained from 10 pairs of training and test sets revealed that the proposed reduction methodology is a robust one and it also reduces noise in data which in turn improves the prediction outcomes. The average percentage measures of: (i) the correct prediction increases by 26%, (ii) the false positive decreases by 36%, (iii) the false negative decreases by 89%, and (iv) the unpredictable objects increases by 136% which is the indicative of a reliable system. Prediction of no decision for an object is always preferred over prediction of a false positive or a false negative decision. The neighborhood-based reduction system also increases the granularity of the dataset which is different from the increase in the granularity through the use of a generalization process.**

Keywords—Data Mining; Big Data; Data Volume Reduction; Neighborhood System; Property Preservation; Organic Discretization.

I. INTRODUCTION

A very large dataset may be mined for the purpose of association analysis, concept analysis, decision support analysis, market analysis, and prediction, to name a few. The sheer volume of a very large dataset is the major obstacle in mining the data because the size of the dataset is beyond handling abilities of the traditional methodologies. Any methodology used for reducing the size of the dataset must be able to preserve the *intrinsic properties* of the very large dataset; otherwise, the methodology is not a robust one.

To remove, or at least ease, the volume obstacle, partitioning methodologies have been contemplated [1]. In any partition-based methodology, the very large dataset is divided into partitions either randomly or based on some criteria suggested by the mining goal. The mining of each partition takes place separately. However, the mining outcome (intrinsic properties) of a very large dataset is not equivalent of the union of the intrinsic properties of the individual partitions. Reader needs to know that the parallel processing plays a big role in mining of very large datasets and datasets are segmented for use by the parallel processor [2]. This segmentation is different from partitioning because during the segmentation process the dataset is perceived as one entity, whereas the partitioning process perceives each partition as a separate dataset.

Clustering-based methodologies may also reduce the volume of data [3]. The common practice is that a cluster of records of the very large dataset is replaced by the seed of the cluster. The inclusion of a record in a cluster is based on the fact that the sum of its attribute distances from the corresponding attributes of the seed is less than a threshold distance. The problem with clustering is that it is influenced by the sum of the individual attribute's differences and not by the differences of the individual attributes. As a result, a cluster satisfies a condition that does not guaranty the true homogeneity of its record members. Replacing a cluster of non-homogenous records with its seed has a dire effect on the preservation of the properties of the large dataset.

In this paper, we propose a methodology, *neighborhood system*, for volume reduction of very large datasets. We also empirically show that the reduction methodology is a

 ISBN: 978-1-61208-358-2

robust one. That is, the reduced dataset preserves the intrinsic properties of the original dataset.

The organization for the rest of the paper is as follow: The previous work is presented in Section two. The methodology is the subject of Section three. The empirical results are covered in Section four. The conclusion and future research are discussed in Section five.

II. PREVIOUS WORKS

The data reduction has been explored by researcher for four different purposes of *data storage*, *data transmission*, *data presentation* and *data mining*. For data storage, basically, data is compressed to take less space. A compressed dataset may be decompressed as needed. Numerous data compression techniques have been reported in literature [4][5]. Some compression techniques are lossy and some are lossless. Use of lossy techniques compresses data in a way that it cannot turn completely into its original form upon applying the decompression process. In contrast, data compressed by using the lossless compression techniques turned into the original dataset after decompression.

For data transmission, data is reduced during its preparation for the transmission and usually returns to its original form at the destination. For example, prior to transmission of an image through a communication channel, the image is reduced to lower the communication time and be adapted to the communication channel limitations [6][7].

For data presentation, data is reduced using different way of its presentation. For example, visualization of data presents data in a reduced form [8][9]. As another example, collection of a high volume of raw data is used for building a product. By doing so, the final product becomes the reduced version of the raw data [10][11].

For data mining, data is reduced horizontally and vertically prior to applying any data mining methodology. The horizontal data reduction means removing the redundant attributes from a dataset. Entropy analysis, correlation analysis, relevancy analysis, and rough sets are some of the well-known methods for performing the horizontal reduction [12][13][14][15]. The vertical reduction reduces the number of records in a dataset. This is done through collapsing the duplicated records and in some cases removal of conflicting records. Such reduction is a part of the pre-mining process and the reduced datasets often have slightly less number of records than the original datasets. For very large datasets, a considerable vertical reduction in addition to the vertical reduction prescribed by the pre-mining process is needed. However, the reduced version of the dataset ought to preserve the intrinsic properties of the original dataset; otherwise, it is a useless reduction.

In this paper, we propose and investigate a robust vertical reduction methodology that is able to (a) reduce the size of dataset beyond pre-mining reduction and (b) preserve the intrinsic properties of the original dataset. To the best of our knowledge, there is no such investigation reported in the literature.

III. METHODOLOGY

We present, first, the neighborhood system as a new methodology for reducing the volume of a dataset. Second, we introduce formal definition of the *intrinsic properties* (or simply *properties*) of a dataset along with the methodology for testing the property preservation. Finally, we present the organic discretization in support of property preservation.

A. The Neighborhood System

A dataset is a collection of records and each record has n attributes $U = \{A_1, \ldots, A_n\}$. Consider records R_i: $(v_1, \ldots, v_n)$ and R_j: $(v'_1, \ldots, v'_n)$, (values of v_k and v'_k belong to attribute A_k). R_j is the *neighbor* of R_i in reference to U, if $|v_i - v'_i| \leq r$ (for $1 \leq i \leq n$). r is a radius threshold. All the neighbors of R_i within a given dataset make the *neighborhood* of R_i. It is true to say that every record is also a member of its own neighborhood. The following notation is used to denote the neighborhood of R_i: $N(R_i)_{[U, r]}$. If R_j is in $N(R_i)_{[u,r]}$, then R_i is also in $N(R_j)_{[u,r]}$.

Since the threshold radius can take many different values, the record R_i may have many, not necessarily distinct, neighborhoods. This is true for all the records in the dataset. The neighborhoods of every record of a dataset are collectively referred to as a *neighborhood system* of the dataset.

Hashemi et al. [16] divide the neighborhood system for each record into three regions of *closest*, *closer*, and *close* neighborhoods. These regions for R_i are defined as Closest $(R_i) = N(R_i)_{[U, r=0]}$, Closer $(R_i) = N(R_i)_{[U, r=a]}$, and Close $(R_i) = N(R_i)_{[U, r=b, \text{ where } b>a]}$. The three regions are also known as *the workable neighborhoods* of R_i.

TABLE I: A DATASET.

Records	A1	A2	A3	A4	A5
R_1	1	2	1	3	4
R_2	1	1	2	2	2
R_3	2	2	3	1	2
R_4	1	2	1	3	4
R_5	2	3	2	2	3
R_6	3	1	3	1	2
R_7	2	1	1	2	3
R_8	3	2	2	3	3

As an example, consider the dataset of Table 1. For the record R1, the workable neighborhoods are:

$\text{Closest}(R_1) = \{R_1, R_4\}$

$\text{Closer}(R_1) = \{R_1, R_4, R_5, R_7\}$

$\text{Close}(R_1) = \{R_1, R_2, R_4, R_3, R_7, R_5, R_6, R_8\}$

For identifying the closer and close neighborhoods of R_1, we use a = 1 and b = 2. Therefore, the closest, closer, and close neighborhoods of R_1 include those records of the dataset that their attribute values differ from their

 ISBN: 978-1-61208-358-2

corresponding attribute values in R_1 by zero, absolute value of one, and absolute value of 2, respectively.

Since, in this research effort, the goal of the mining of a very large dataset is to perform "prediction", we assume that each record has an extra attribute of *decision*, Table 2.

The decision attribute does not play any role in finding a neighborhood. However, the decision attribute is used to assign a *certainty factor* to any neighborhood of interest. For example, the certainty factor for the Closer(R_i) is α = N/|Closer(R_i)|, where N is the number of records in the Closer(R_i) who have the same decision value as R_i. As a result, the certainty factor for the Closest(R_1), Closer(R_1), Close(R_1) are 1, 1, and 5/8, respectively.

TABLE II: A DATASET WITH A DECISION ATTRIBUTE.

Records	A1	A2	A3	A4	A5	Decision
R_1	1	2	1	3	4	1
R_2	1	1	2	2	2	0
R_3	2	2	3	1	2	0
R_4	1	2	1	3	4	1
R_5	2	3	2	2	3	1
R_6	3	1	3	1	2	1
R_7	2	1	1	2	3	1
R_8	3	2	2	3	3	0

1) Record Tree

Let us focus on the record R_i and its closer neighborhood. Initially, the *record tree* of the R_i is the presentation of Closer(R_i) in form of a tree for which R_i is the root and the neighbors of R_i are the children. The tree is assigned a *certainty factor* and a *signature*. The certainty factor of the tree is the same as the certainty factor of the Closer(R_i). The signature of the tree is a record with the same number of attributes as R_i and the value for attribute A_m of the signature is the average of values of attribute A_m for all the records in the record tree.

Each child, C_i, of the tree is expanded as a new sub-tree by the records in Closer(C_i). The new sub-tree is pruned based on the following criteria:

a. If a child of C_i is already appeared as a node somewhere in the tree, the child is pruned.
b. If the Euclidean distance of a child of C_i from the signature of the tree is greater than a given threshold value, the child is also pruned.

After the expansion of Ci, if any of its children survived the pruning process, the record tree of R_i becomes a new tree with a new certainty factor and a new signature. The process of expansion of the new record tree for R_i continues in a breadth-first fashion until it cannot be expanded any longer. All the records that are part of the totally expanded record tree of R_i will not have their own record trees and cannot be a part of another record tree. However, the building of the record trees for the remaining records of the dataset is a continual process.

Selection of R_i for building its record tree is not a random act. R_i is selected such that its closer neighborhood has the highest certainty factor among all the closer neighborhoods of the dataset. In case of a tie, R_i has the highest cardinality. If having a tie persists, R_i is selected randomly among the qualified records.

The signature of the record tree for Closer(R_i) acts as a representative of all the records in the neighborhood and replaces all of them. Let us assume that the process of building record tree for the records of a dataset produces T record trees. The ratio of |dataset|/T is the *reduction factor*. For the dataset in Table 2 and for the Euclidean distance threshold of 3.7, only two record trees are produced: see Figure 1. Therefore, the reduction factor is 8/2 = 4.

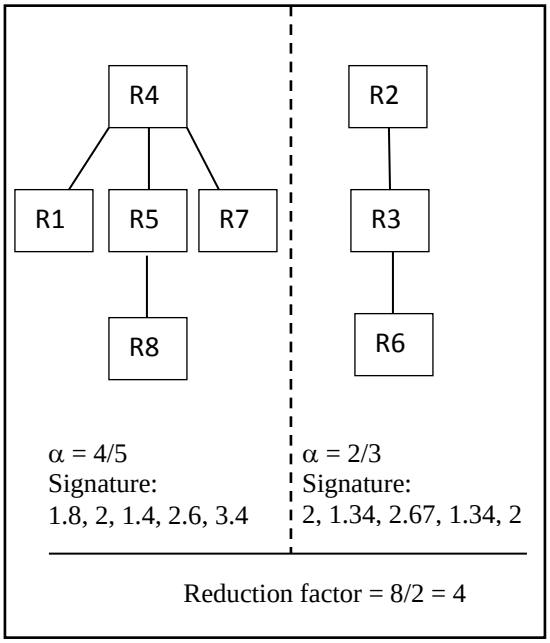


Figure 1. Record trees for the dataset of Table 2 along with their certainty factors and signatures.

B. Properties of a Dataset

So far, the proposed neighborhood system is able to reduce a given very large dataset, V, by factor of K into a new dataset, V'. In this section, we define the intrinsic properties and describe the methodology for checking the preservation of the properties of V by V'.

The properties of a very large data set are a sextuple, (V, G, A, F, Q, E), where:

V is a very large dataset,
G is the goal of the mining process
A is a methodology used for reaching the goal of G.
F is resulting set of findings applying A on V,
Q is the quality measure for F. The quality of F is usually measured by using another dataset (E).
E is the entity involved in measurement of the quality for F.

The quality measure of F needs an explanation. Let us assume G is "prediction" and A is the "ID3" algorithm [17]. The outcome of applying A on V is a set of prediction rules, F. The quality measure for F is the quality of the prediction for the test records using F. Thus, the test set is E.

As another example, let us assume G is "basket analysis" and A is the "Apriori" algorithm [18]. The outcome of applying A on V is a set of frequent itemsets, F. The quality measure of F is the collection of the quality measures for each frequent itemset which is presented in pair of (*support, confidence*). Calculation of support and confidence are done using records in V. Therefore, V is the entity involved in quality measure of F.

Definition: Let V' be a reduced version of V and the mining goal for both V and V' be the same. Let also A be a well-established algorithm for reaching the goal. In addition, let F and F' be the two sets of findings produced by applying A on V and V'. In addition, let Q and Q' be the quality measures of F and F', respectively, calculated using the same entity E. The properties of V and V' are: *(V, G, A, F, Q, E)* and *(V', G, A, F', Q', E)*. If Q' = Q, then the properties of V has been preserved by V'. If Q' > Q then V' has not only preserved the properties of V but also reduced noise in data.

Let us assume that the algorithm A cannot be applied on V' due to the fact that data in V' is continuous, whereas data in V is discretized (required by the algorithm A). To remove this obstacle, either data in V' needs to be discretized or algorithm A needs to be replaced by another algorithm that can process both discrete and continuous data. The first option is more logical because it does not limit the list of algorithms that can be applied on V'. As a result, we introduce our own discretization methodology named *organic discretization* in the following sub-section.

C. Organic Discretization

The majority of the discretization methodologies, reported in literature, have an artificial discretization theme [19][20]. For example, the interval between the maximum and minimum values of an attribute is divided into a number of equal width smaller intervals and each small interval is assigned a discrete value that replaces all the values within the small interval. Such discretization is artificial and does not consider any characteristics of the values of the attribute. Although some of the methodologies such as bin-based and radius-based try to ease the problem, but they cannot avoid artificially discretizing the data [2]. There are more sophisticated discretization methodologies that are so labor intensive that their use is not cost effective [2].

In this section, we propose an organic discretization methodology that uses the closeness of values in an attribute for discretization. The methodology is simple and organic. Because of that the intervals represented by discrete values do not have necessarily the same width and the width of each interval is decided by the data itself.

In this methodology, the values of the attribute are sorted in ascending order and the differences between every two adjacent values are measured. A user selects a preferred small difference, P_d. The end point of a current interval is decided based on the differences between the value located in locations L and L+1 in the list of sorted values and P_d. If P_d is zero, then every unique value belongs to a new interval. If P_d is too large, then the entire attribute becomes one interval. The best choice for P_d to discretize an attribute of dataset V' is to generate the same or close number of intervals—and therefore, discrete values—for the attribute as there is for the corresponding attribute in V. The reason stems from the fact that some mining algorithms choose attributes with more discrete values over those with less number of attributes or vice versa. Since the goal is to investigate the preservation of the properties in reduced datasets, we want to remove any biases causing by the number of discrete values.

The following algorithm provides the details of the organic discretization approach:

Algorithm Organic

Input: A dataset with n attributes of A1, . . ., An. Data of the dataset is continuous. Two threshold values of t_d and t_{count}.

Output: The discretized dataset.

Step1. Repeat for each attribute A_i

Step2. B = Ai, sorted in ascending order.

Step3. C[i] = ABS(B[i]-B[i+1]).

Step4. Locate in C those elements with value > t_d and Collect their indices in array D.

Step5. top = 1; bottom = 1; Count = 0; //Top and bottom are pointers pointing to the first element of interest in array B and first element of interest in D;

Step6. Repeat Steps 7 to 9 while top < |B|;

Step7. If D= ∅, then bottom = |B|;

Step8. Those values in array B from B[top] to B[D[bottom]] make an interval, Int, represented by a discrete value which is the median of the values in the interval; count++;

Step9. top = top + |inter|; Remove the first element of D;

Step10. If count > t_{count}, then increase t_{count}; go to Step2;

Step11. End;

The variation of the algorithm may be considered by the choosing a different value to represent the interval produced in Step 8. We used the median value to represent the interval.

IV. EMPIRICAL RESULTS

In a glance, we: (i) generate 10 pairs of the training and test sets out of the original dataset, (ii) generate the reduced version of the same 10 training sets using the neighborhood

 ISBN: 978-1-61208-358-2

system and the organic discretization approaches, (iii) select a mining goal and a well-known algorithm to achieve it. If the average results produced by applying the well-known algorithm on the second ten pairs (the reduced ones) is the same or better than the average results produced by applying the algorithm on the first ten pairs (the original ones), then the reduced version of the training sets has preserved the intrinsic properties of the original dataset; Otherwise, the intrinsic properties have been damaged and the reduction methodology is not robust.

To provide the details of the process, the "prediction" is our mining goal and the algorithm to achieve the goal is the well-known ID3 algorithm. We have a dataset with 1000 records and each record has 8 attributes. Each one of the first seven attributes has six possible discrete values. The last attribute is a decision and it has two possible values of 1 and 0. One may point out that the dataset is not a very large dataset. However, here our goal is to show the proof of concept.

Ten percent of the records with decision 1 and ten percent of the records with decision zero have been set aside to make one test set. Among the remaining m records, m1 of them has decision one and m2 of them has decision zero (m = m1 + m2 and m2 <m1). We pick m2 records out of the m1 records with decision one along with the entire m2 records with decision zero and make the training set.

By repeating the same process, we generated 10 pairs of the training (Tr) and test (Ts) sets such that $Ts_i \cap Ts_j = \emptyset$ (for i = 1 to 10, j = 1 to 10, and i ≠ j) and $Tr_i \cap Ts_i = \emptyset$ (for i = 1 to 10). Since the original dataset is made up of the discrete values, so the training and test set pairs.

We have also generated:

1. A reduced version of each training set by applying the neighborhood methodology on the set (the reduction factor of the training sets was varying from 3.2 to 4.55 for different training sets). As a result, we produced 10 reduced training sets. The data of the reduced training sets were no longer discrete values.
2. A discretized version of each reduced training set by applying the organic discretization methodology on the set. It was clear that the new discretized values in the training set did not have the same meaning as the discrete values in the corresponding test set. Therefore, the discretization intervals of data established by the organic discretization of the training set were used to discretize the original test set.

To sum-up, we ended up having 10 pairs of the training and test sets build out of the original dataset and 10 pairs of the same training and test sets with the new discrete values influenced by the neighborhood methodology. The first and the second 10 pairs are referred to as the *Original* and *Reduced* sets, respectively.

To investigate the preservation of the properties, we took the following step for each training and test pair in the Original and Reduced sets:

- ID3 was applied on the training set and the prediction rules were obtained and used to predict the decision for the records of the test set. The quality of the prediction was measured by calculating the percentage of the number of correct predictions, false positives, false negatives, and not predictable records. The quality of the prediction for the Original pairs and Reduced pairs are shown in Table 3 and Table 4, respectively.

TABLE III: THE QUALITY MEASURES OF THE PREDICTION PROCESS FOR THE ORIGINAL SET USING ID3.

Original Pairs	% Correct Predictions	% False (+)	% False (-)	% Not-predictable
1	47.6	48	5	0
2	59.5	29	11	0
3	64.3	34	2	0
4	57.2	31	11	0
5	40.5	52	7	0
6	61.9	24	11	2
7	78.6	12	9	0
8	47.6	41	11	0
9	54.8	26	11	7
10	52.4	36	9	2
Avg.	56.4	33.3	9.7	1.1

TABLE IV: THE QUALITY MEASURES OF THE PREDICTION PROCESS FOR THE REDUCED SET USING ID3.

Reduced Pairs	% Correct prediction	% False (+)	% False (-)	% Not-predictable
1	66.7	24	0	10
2	76.2	19	0	5
3	66.7	26	1	7
4	71.4	17	1	11
5	71.4	21	1	7
6	76.2	12	2	10
7	69	14	0	17
8	66.7	24	0	10
9	64.3	12	1	23
10	73.8	12	2	12
Avg.	70.24	18.1	0.8	11.2

Since the average performance of ID3 on the Reduced set is much better than the average performance of ID3 on the Original set, the intrinsic properties of each test set has been preserved.

One may raise the following question: Is the property preservation possible using a prediction algorithm other than ID3? To answer this question we also conducted the same experiment using the Rough Sets algorithm [15][21][22][23]. The results for the Original and Reduced sets are shown in Table 5 and Table 6, respectively.

 ISBN: 978-1-61208-358-2

The Average prediction performance of the Rough Sets approach on the Original and the Reduced sets support the findings through the use of ID3.

V. CONCLUSION AND FUTURE RESEARCH

The results of Tables 3, 4, 5, and 6 reveal that the proposed reduction methodology preserves the intrinsic properties of the original dataset. Considering all four tables, the average percentage measure of: (i) the correct prediction increases by 26%, (ii) the false positive decreases by 36%, (iii) the false negative decreases by 89%, and (iv) the unpredictable records increases by 136% which it is the indicative of a reliable system. Prediction of a "no decision" for an object is always preferred over prediction of a false positive or a false negative decision.

TABLE V: THE QUALITY MEASURES OF THE PREDICTION PROCESS USING ORIGINAL SET AND ROUGH SETS.

Original Pairs	% Correct Predictions	% False (+)	% False (-)	% Not-predictable
1	40.2	38	11	10
2	53.1	22	16	9
3	60.8	14	15	10
4	41.7	27	11	10
5	40.4	38	9	12
6	55.7	24	12	8
7	62.9	12	15	10
8	75.8	12	7	5
9	66.8	16	10	7
10	40.1	36	19	5
Avg.	53.75	23.9	12.5	8.6

TABLE VI: THE QUALITY MEASURES OF THE PREDICTION PROCESS USING REDUCED SET AND ROUGH SETS.

Reduced Pairs	% Correct prediction	% False (+)	% False (-)	% Not-predictable
1	68.3	18	5	9
2	72.9	19	2	7
3	62.2	28	1	9
4	70.8	17	2	10
5	68.2	24	2	5
6	75.1	13	2	10
7	69	14	1	16
8	63.1	26	1	9
9	64.3	13	0	23
10	67.8	14	0	19
Avg.	68.17	18.6	1.6	11.7

To explain an interesting observation, let us briefly talk about noisy data which is a synonym for the erroneous data [1]. Error (noise) in data is resulting from corruption of data at the time of collection and or inputting. The noise in data is considered as an obstruction in any data mining process including prediction. The improvement of the prediction results for the Reduced set, by both ID3 and Rough Sets algorithms, indicates the fact that the data in the Reduced set has less noise than data in the Original set. Therefore, the proposed data reduction methodology not only preserves the intrinsic properties of the Original set but it also decreases the noise in the set.

The neighborhood-based reduction system also increases the granularity of the dataset which is different from the increase in the granularity through the use of a generalization process. To explain it further, let us assume that a dataset contains the monthly profit reported for a given company for duration of one year. This dataset has 12 records (one per month). One may add up the monthly profits for each quarter to express the quarterly profit. In this case, the dataset is reduced and it has only four records. The number of records may change into only 2 records, if bi-annual profits is sought. The reduction of the records provides different granules in each case. In the first and the second reductions each granule represents quarter profits and bi-annual profits, respectively. The reductions are also known as the representations of the profits for two *foot-steps* ("quarter" and "bi-annual") within the *concept hierarchy* of the time.

The prediction rules obtained from the higher granules may not preserve the properties of the dataset. The reason stems from the fact that (i) a higher granule ignores the details of lower granules and (ii) the foot-steps in a concept hierarchy are natural steps within the domain of the interest (in our example time domain) and does not have anything to do with the closeness of values of the records' attributes within the foot-step.

In contrast, the granularity provided by the proposed reduction methodology is only based on the closeness of values of the records' attributes. The foot-step based granularity still can be applied to the granules delivered by the proposed reduction system.

One of the challenges in this research effort was the selection of a representative for the interval produced in Step 8 of the Algorithm Organic. On the whole, there are seven possible options; thus, seven variations of the algorithm may be used. The options are: (i) the median value within the interval when the number of records, n, in the interval is odd, (ii) the $(n/2)^{th}$ value when n is even, (iii) the $[(n/2) +1]^{th}$ value when the n is even, (iv) average of $(n/2)^{th}$ value and $[(n/2)+1]^{th}$ value when n is even, (v) the first number in the interval, (vi) the last number in the interval, and (vii) average of all the values in the interval. We have chosen options (i) and option (ii) for the cases that n is odd and even, respectively. The methodology used for selecting these two options was the "trial-and-error" approach.

It was also noted that the robust reduction methodology for mining the prediction rules, may not be able to preserve the properties of the dataset for another mining goal –say, association analysis. For example, let us assume that we are interested in learning about the correlation between two values of "a" and "b" that belong to two different attributes

 ISBN: 978-1-61208-358-2

of a given dataset. The reduction process may change the value of "a" into possibly m new and different values. The value "b" may also be changed into possibly k new and different values (m and k are not necessarily equal). As a result, finding the correlation between values of "a" and "b" within the reduced dataset may be a moot point. However, one may argue that the correlation between "a" and "b" may be preserved within the discretized values produced through application of the Algorithm Organic on the reduced dataset. Such possibility is under investigation to determine whether or not the preservation of properties by a reduced dataset is sensitive to the purpose (goal) of mining.

In addition, the application of the proposed methodology on a very large dataset is under investigation which includes the viability study of the signatures as prediction rules along with the horizontal and vertical reductions of the signatures.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann publishers, 2001.

[2] M. Vojnovi´c, F. Xu, and J. Zhou, "Sampling Based Range Partition Methods for Big Data Analytics," Technical Report of MSR-TR-2012-18, Microsoft Corporation, Redmond, WA, March 2012.

[3] M., Meilă, "Comparing Clusterings by the Variation of Information," Learning Theory and Kernel Machines, B. Schölkopf and M. K. Warmuth (Eds.), Springer Lecture Notes in Computer Science, Volume 2777, 2003, pp. 173–187

[4] T. Bell, I. H. Witten, and J. G. Cleary, "Modeling for text compression," The ACM Journal of Computing Surveys, vol. 21, no. 4, Dec. 1989, pp.557-591.

[5] S. Mahmud, "An Improved Data Compression Method for General Data", International Journal of Scientific & Engineering Research, vol. 3, no. 3, March 2013, pp.1-4.

[6] J. H. Pujar and L. M. Kadlaskar, "A new Lossless Method of Image Compression and Decompression Using Hufman Coding Techniques," Journal of Theoretical and applied Information Technology, vol. 15, no. 1, May 2010, pp. 18-23.

[7] D. Taubman, "High performance scalable image compression with EBCOT," IEEE Transactions on Image Processing, vol. 9, no. 7, July 2000, pp.1158-1170.

[8] Y. S. Wang, C. Wang, T. Y. Lee, and K. L. Ma, "Feature-preserving volume data reduction and focus+context visualization," IEEE Transaction: Visualization and Computer Graphics, vol. 17, no. 2, Feb. 2011, pp. 171-181.

[9] E. R. Tufte, "The Visual Display of Quantitative Information," 2nd Ed, Graphic Press Publisher, May 2001.

[10] S. Kang, J. Lee, H. C. Kang, J. Shin, and Y. G. Shin, "Feature-preserving reduction of industrial volume data using gray level co-occurrence matrix texture analysis and mass-spring model," Journal of Electronic Imaging, vol. 23, no. 1, Feb. 2014, pp. 13-22.

[11] T.R. Jones, "Feature Preserving Smoothing of 3D Surface Scans," MS. Thesis, Computer Science Department, MIT, Sept. 2003.

[12] R. M. Gary, "Entropy and Information Theory," Springer, 1990.

[13] J. Natt, R. Hashemi, A. Bahrami, M. Bahar, N. Tyler, and J. Hodgson, "Predicting Future Climate Using Algae Sedimentation," The Ninth International Conference on Information Technology: New Generation (ITNG-2012), Sponsored by IEEE Computer Society, Las Vegas, Nevada, April 2012, pp. 560-565.

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

[15] R. Hashemi, A. Tyler, and A. Bahrami, "Use of Rough Sets as a Data Mining Tool for Experimental Bio-Data," Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications, T. G. Smolinski, M. G. Milanova, and A. E. Hassanien, (Eds.), Springer-Verlag Publisher, June 2008, pp. 69-91.

[16] R. Hashemi, A. Bahrami, M. Smith, N. Tyler, M. Antonelli, and S. Clapp, "Discovery of Predictive Neighborly Rules from Neighborhood Systems," International Conference on Information and Knowledge Engineering (IKE'13), Las Vegas, Nevada, July 2013, pp. 119-125.

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

[18] R. Hashemi, L. LeBlanc, and B. Westgeest, "The Effects of Business Rules on the Transactional Association Analysis," The 2004 International Conference on Information Technology: Coding and Computing (ITCC-2004), Pradip K. Srimani (Editor), Sponsored by IEEE, Las Vegas, Nevada, vol. II, April 2004, pp. 198 - 202.

[19] M. R. Chmielewski, and J. W. Grzymala-Bussse, "Global discretization of continuous attributes as preprocessing for machine learning, vol. 15, no. 4, Nov. 1996, pp. 319-331.

[20] J. Zhao and Y. H. Zhou, "New heuristic method for data discretization based on rough set theory," The journal of China Universities of Posts and Telecommunications, vol. 16, no. 6, Dec. 2009, pp. 113-120.

[21] Z. Pawlak, J. W. Grzymala-Busse, R. Slowinski, and W. Ziarko, "Rough Sets", Communications of ACM, vol 38, no. 11, Nov. 1995, pp. 88-95.

[22] R. Hashemi, B. Pearce, R. Arani, W. Hinson, and M. Paule. "A Fusion of Rough Sets, Modified Rough Sets, & Genetic Algorithms for Hybrid Diagnostic Systems", In: Lin TY, Cercone N Editors. Rough Sets & Data Mining: Analysis of Imprecise Data. Kluwer Academic Publishers, 1997. pp. 149-176.

[23] R. Hashemi, F. Choobineh, W. Slikker, and M. Paule, "A Rough-Fuzzy Classifier for Database Mining", The International Journal of Smart Engineering System Design, no. 4, 2002, pp. 107-114.

 ISBN: 978-1-61208-358-2

A Decision Support Approach for Quality Management Based on Artificial Intelligence Applications

Nafissa Yussupova, Maxim Boyko, Diana Bogdanova
Faculty of informatics and robotics
Ufa State Aviation Technical University
Ufa, Russian Federation
Emails: {yussupova@ugatu.ac.ru, maxim.boyko87@gmail.com, dianochka7bog@mail.ru}

Andreas Hilbert
Faculty of Economics
Dresden University of Technology
Dresden, Germany
andreas.hilbert@tu-dresden.de

***Abstract*—This paper describes the application of a novel domain-independent decision support approach for product quality management. It is based on customer satisfaction research through deep analysis of consumer reviews posted on the Internet in natural language. Artificial Intelligence (AI) techniques, such as Text Mining and Data Mining, are used for realization of consumer reviews analysis. In paper, specific Internet resources (such as yelp.com, tripadvisor.com, tophotels.ru) are used for accumulating customer reviews as a data source is considered. This is performed in accordance with the quality standard ISO 10004 proposed decision support approach, which allows for both qualitative and quantitative customer satisfaction surveys to be carried out. The output of the quantitative survey are values of customer satisfaction by product and each product's aspect. The output of the qualitative survey are significance values of products aspect for customers and identified latent relations between satisfaction by product and satisfaction by products' aspects. The proposed approach is performed as a prototype of a Decision Support System. To evaluate the efficacy of the proposed approach, an experiment on hotel's customer satisfaction has been carried out. The obtained results prove the efficacy of the proposed decision support approach for quality management and the concept of using it instead of classical methods of qualitative and quantitative research of customer satisfaction.**

Keywords-quality management; decision support system; sentiment analysis.

I. INTRODUCTION

In order to provide product quality, a company should make effective managerial decisions. In the modern world, the efficacy of managerial decision-making process depends on the information available to the person that makes decisions and the depth of information analysis. Therefore, a company should develop processes of automated collection of information and its further processing and analysis. Decision-making should be based on the knowledge and principles obtained during the analysis of the collected data.

Quality assurance is currently attained through a process approach based on the model of a quality management system [1]. It describes the interaction of the company and the customer during the process of product production and consumption. To correct the parameters of a product's quality in order to improve it for the customer, the model has feedback. For companies, feedback during the process of quality management is the information about the level of customer satisfaction, which is expressed in the form of customer reviews about a product's quality. That is why customer satisfaction is key information for quality management that influences decision-making.

To collect data and evaluate customer satisfaction, International Quality Standards ISO 10004 (International Organization for Standardization) recommends using the following classical methods: face to face interviews, telephone interviews, discussion groups, mail surveys (postal questionnaires), on-line research and surveys (questionnaire surveys) [2]. However, these methods of collection and analysis of customer opinions have a number of significant drawbacks.

A general drawback of these methods is a large amount of manual work: preparing questions, creating a respondent database, mailing questionnaires and collecting results, conducting a personal interview, and preparing a report. All of these procedures make a research expensive. These methods cannot monitor customer satisfaction continuously. For this reason, monitoring is limited by a one time period. There is no possibility for monitoring trends of customer satisfaction. It also has a negative influence on lengthiness of managerial decision making.

Another problem regards various scales for measuring customer satisfaction and their subjectivity perception. Value of customer satisfaction is estimated by abstract satisfaction indices that are difficult to understand, hard to compare and interpret. Furthermore, methods for data analysis recommended by ISO 10004 [2] allow detection of only linear dependencies and relations in data.

 ISBN: 978-1-61208-358-2

The aim of this paper is the development of a decision support approach for quality management based on the research of customer satisfaction with use of AI technologies.

The remainder of this paper is organized as follows: in Section 2, we focused on overview of recent solutions and frameworks for analysis of user generated content and their drawbacks. In Section 3, we described architecture and workflow of proposed decision support system. In Section 4, we described using AI techniques for qualitative and quantitative customer satisfaction surveys. In Section 5, we provide experiment with researching customer satisfaction of two hotels and whole resort. The obtained results could be used for decision making.

II. RELATED WORK

Applying Text Mining tools for analyzing customers' reviews posted on the Internet is not novel. There are many studies concerning models and methods for data collection, sentiment analysis and information extraction. Recent studies show acceptable accuracy of methods for sentiment classification. Gräbner et al. [3] proposed a system that performs the sentiment classification of customer reviews on hotels. Lexicon-based method [27] allowed the correct classification of reviews with a probability of about 90%. These achievements make sentiment analysis applicable for an application on quality management.

Jo and Oh [4] and Lu et al. [5] considered the problems of automatically discovering products' aspects and sentiments estimation for these aspects which are evaluated in reviews. For solving these problems, they suggested methods based on Latent Dirichlet Allocation [28] and its modifications.

A lot of social monitoring systems and frameworks have been developed for automatic analysis of reviews and topics. Liu et al. [6] presented framework called Opinion Observer for analyzing and comparing consumer opinions of competing products. This prototype system is able to visualize the strengths and weaknesses of each product in terms of various product features. Kasper and Vela [7] presented a web based opinion mining system for hotel reviews and user comments that supports the hotel management called BESAHOT. The system is capable of detecting and retrieving reviews on the web, classifying and analyzing them, as well as generating comprehensive overviews of these comments. Blair-Goldensohn et al. [8] proposed a system that summarizes the sentiment of reviews for a local service, such as a restaurant or hotel. In particular, they focus on aspect-based summarization models. Ajmera et al. [9] developed a Social Customer Relationship Management (SCRM) system that mines conversations on social platforms to identify and prioritize those posts and messages that are relevant to enterprises. Bank [13] proposed interactive Social Media monitoring system to extract related information from user generated content. One of the important contribution of this work was the proposition of new quality indeces.

In some related work, authors pay attention to relations between overall ratings of products, and ratings of products' aspects evaluated in the review. Wang et al. [10] formulated a novel text mining problem called Latent Rating Analysis (LARA). LARA aims at analyzing opinions expressed in each review at the level of topical aspects to discover each individual reviewer's latent rating on each aspect as well as the relative importance weighted on different aspects when forming the overall judgment. De Albornoz et al. [11] aimed to predict the overall rating of a product review based on the user opinion about the different product features that are evaluated in the review. For experiments, authors used reviews on hotels.

Wachsmuth et al. [12] formulated and validated an important hypothesis that the global sentiment score of a hotel review correlates with the ratio of positive and negative opinions in the review's text and that the global sentiment score of a hotel review correlates with the polarity of opinions on certain product features in the review's text.

The main drawback of these considered systems is that they can provide entirely only a quantitative survey of customer reviews, i.e., provide measurement of the degree of customer satisfaction by a product and its aspects. Qualitative survey were usually only conducting the extraction of a product's aspects. Estimation of the significance of each product's aspects for the customer is missed. The information about products' aspects that influence satisfaction and their relative importance for the customer is missing, as well as an insight into customer expectations and perceptions.

The most related work to this problem is [24]. It is dedicated to the topic of aspect ranking, which aims to automatically identify important aspects of product from online consumer reviews. Most proposals used a probabilistic model with a large number of parameters that lead to low robustness of the model. Total weighting values of aspects are calculated as the average of the weighting values by each review. Finally, significance values of aspects are estimated independently of an opinion's sentiment, e.g., in real life, we can discuss in review about bad "signal connection", but we usually omit comments in case of good "signal connection", because it must be in phone. In this manner, it is possible to use the Kano's model of customer satisfaction [25], which classifies customer preferences into four categories.

In this paper, for qualitative survey is used a novel approach based on transformation results of Sentiment Analysis into binary data. After that, binary data is processed with a Data Mining tool – Decision Tree (see Section IV). Qualitative survey aims to identify how the sentiment of reviews depends on the sentiment of different products' aspects. In other words, how overall customer satisfaction by product depends on the customer satisfaction by a product's aspects. Decision Tree performs this aim and identifies latent relations between the sentiment of reviews and sentiment of a product's aspects. We estimate the significance of aspects from the constructed Decision Tree. Output of qualitative survey are significance values of positive and negative mentions about a product's aspects for customers and identified latent relations extracted by Decision Tree. The availability of both quantitative and qualitative surveys

 ISBN: 978-1-61208-358-2

allows realizing Decision Support System for Quality Management in accordance with quality standard ISO 10004.

III. THE PROPOSED DECISION SUPPORT APPROACH

The suggested approach to decision making in product quality management accomplished through unification of methods for collecting and processing text data into Intelligent Decision Support System (IDSS). The architecture (subsystems and contained modules) of the obtained IDSS is presented in Figure 1. The subsystem of monitoring and data collection fills the warehouse with customer reviews and other relevant information. It also supports the actuality of data via automated monitoring of Internet resources and carries out data cleansing. The data storage subsystem provides safe-keeping and integrity of collected reviews and results of data processing. In the subsystem of data analysis are realized methods of sentiment analysis, aspect extraction, aspect sentiment analysis, and decision tree. In subsystem of user interaction is visualized results of analysis.

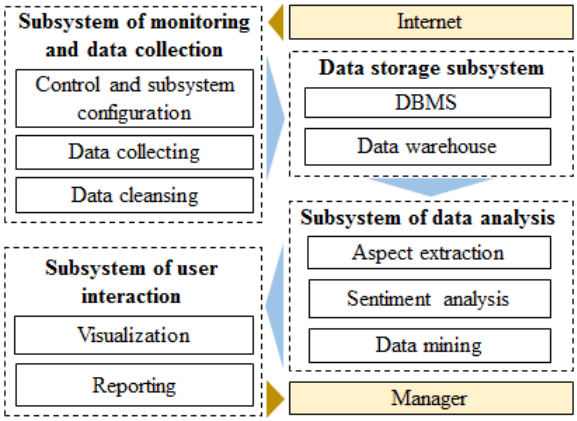


Figure 1. The architecture of Intelligent Decision Support System for product quality management.

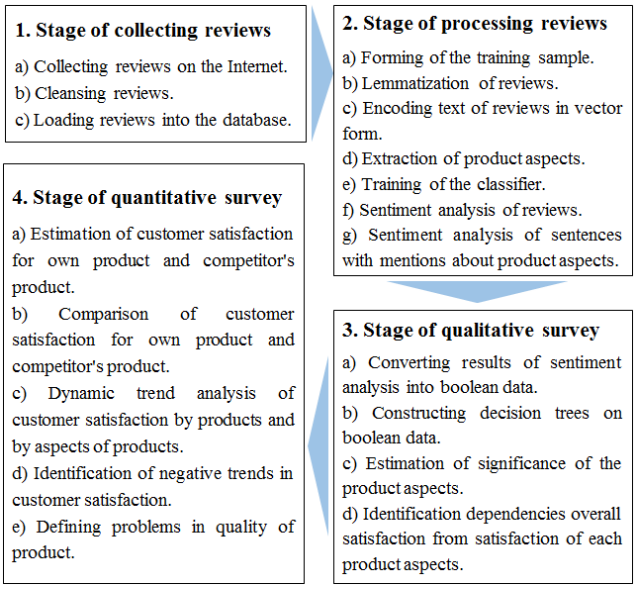


Figure 2. Working algorithm of Intelligent Decision Support System

In Figure 2, the algorithm of the IDSS is presented. It consists of four stages. The first stage includes collection of reviews from Internet resources, data cleansing and loading reviews into the database. The second stage performs processing collected reviews. It includes preprocessing procedures, such as preparing training samples of reviews, text lemmatization, and encoding text of reviews in vector form. Processing procedures include extraction of a product's aspects, training of the classifier and sentiment analysis.

The third stage is the quantitative survey. The quantitative survey is based on sentiment analysis of reviews entirely, and aspect sentiment analysis of sentences with mentions of a product's aspects. Sentiment classification is attained through binary scale – positive and negative sentiments. As a measure of the customer satisfaction is used a ratio of positive reviews (or positive sentences with mentions of a product's aspect) to the sum of positive and negative reviews (or sum of positive and negative sentences with mentions of a product's aspect). The output of the quantitative survey is values of customer satisfaction by a product and each product's aspect.

The fourth stage is the qualitative survey of customer satisfaction. It is based on transformation results of sentiment analysis into binary data and following by constructing of decision tree on it. The qualitative survey aims to identify how sentiment of review depends on the sentiment of different aspects of a product. Decision tree performs this aim and identifies latent relations between sentiment of a review and sentiment of a product's aspects. The output of the qualitative analysis is significance values of a product's aspects for customers and identifying latent relations extracted by decision tree. Managerial decision development and making is carried out on the basis of the performed quantitative and qualitative surveys.

IV. APPLIED ARTIFICIAL INTELLIGENCE TECHNIQUES

In this Section are described implemented AI techniques for customer satisfaction surveys and support decision making.

A. Data collection

Nowadays there are a large number of Internet resources where users can leave their opinions about products and services. The most popular examples of review sites are tophotels.ru (635 thousand reviews), yelp.com (53 million reviews), tripadvisor.com (130 million reviews). Similar resources continue to gain popularity. As opposed to social networking services, the advantage of review sites lies in their purpose - accumulation of customer reviews. One more advantage is that many of such resources have moderators of reviews and confirmation of author's objectivity, e.g., registration procedure.

There are two main types of collecting data from the Internet resources of customer reviews: 1) by using API (Application Programming Interface), and 2) by web data extraction. API is a set of ready-to-use tools - classes, procedures, and functions - provided by the application (Internet resource) for using in an external software product.

 ISBN: 978-1-61208-358-2

Unfortunately, only a few resources that accumulate reviews have API.

In this paper is used the second method for data collection – web data extraction. It is a process of automated content collection from HTML-pages of any Internet resource using special programs or script. Related work is presented in [14][15]. Scheme of reviews collection is presented in Figure 3. Web pages of review sources use HTML (HyperText Markup Language) that sets the structure typical for a review. Such structure includes separate blocks with the name of a product or company with a review, and other blocks with additional information. Therefore, all reviews are clearly identified in relation to the review object. It significantly simplifies the process of data collection in contrast to collecting messages from social networking services.

Input: review site

1. Gathering links on pages with reviews or generating links by using a template.
2. Setting boundaries for a content by using a HTML structure.
3. Collection data on the set of links.
4. Cleaning data and removing duplicates.
5. Loading data into the database.

Output: set of consumer reviews *D*

Figure 3. Algorithm of reviews collection

B. Sentiment Analysis

After data collection, it is possible to process review data with Text Mining tools. In this paper automatic sentiment analysis of reviews is used to evaluate product satisfaction. Sentiment stands for the emotional evaluation of author's opinion about a product that is referred to in the reviews.

There are three main approaches to sentiment analysis: 1) linguistic, 2) statistical, and 3) combined. The linguistic approach is based on using rules and vocabularies of emotionality words [16][17]. This approach is quite time-consuming due to the need of compiling vocabularies, patterns, and making rules for identifying sentiments. However, the main drawback of this approach is the impossibility to get a quantitative evaluation of the sentiment. The statistical approach is based on the methods of supervised and non-supervised machine learning (ML) [18][19]. The combined approach presupposes a combined use of the first two approaches.

In present work the methods of supervised machine learning is used - a Bayesian classifier and Support Vector Machines. Their realization in IDSS is based on techniques described by Pang and Lee [18][19]. More detailed information about implemented methods of sentiment analysis used in this paper can be found in [20][21]. In Figure 4 algorithms of learning and classification for naive Bayes classifier based on Multinomial model are presented. An advantage of these ML methods that they are quite easy in software implementation, and do not require making linguistic analyzers or sentiment vocabularies. They are able to evaluate sentiment quantitatively. For sentiment classification is used binary scale - positive and negative tonality. We use vector representation of review texts with help of the bag-of-words model. As attributes, we consider bit vectors - presence or absence of the word in the review text, and frequency vectors – a number of times that a given word appears in the text of the review. Lemmatization is also used.

Learning of naive Bayes classifier
Input: training set of reviews $D' = \{(d_1, c_1), \ldots, (d_m, c_m)\}$, set of classes $C = \{positive, negative\}$.

1. Extract all words from D' to the vocabulary V
2. For each $c \in C$ do
   3. Count documents N^c in each class c
   4. Calculate probability $p(c) = N^c / N$
   5. For each $w_t \in V$ do
      6. Count number of occurrences $K^c_{w_t}$ of word w_t in each class
      7. Calculate prob. $p(w_t \mid c) = (K^c_t + 1) / \sum_{t=1}^{|V|} (K^c_t + 1)$

Output: V, $p(c)$, $p(w_t \mid c)$

Classification with naive Bayes classifier
Input: review d from set D, V, $p(c)$, $p(w_t \mid c)$

1. Extract all words from d to the vocabulary V_d
2. For each $c \in C$ do
   3. Calculate $score[c] = \ln p(c)$
   4. For each $w_t \in V_d$ calculate $score[c] += \ln p(w_t \mid c)$
5. If $score[pos] > score[neg]$ than $d \in positive$ else $d \in negative$

Output: sentimental label *Sent* of review d (positive/negative)

Figure 4. Algorithm of naive Bayes classifier

C. Aspect Sentiment Analysis

Sentiment Analysis of reviews allows the evaluation of overall customer product satisfaction. However, it does not clearly show what customers like about a product and what they don't like. To answer this question, it is necessary to perform an aspect sentiment analysis. An aspect means characteristics, attributes, and properties that characterize the products, e.g., a "phone battery" or "delivery period". However, one sentiment object can have a great number of aspects. Furthermore, aspects in the text can be expressed by words-synonyms, e.g., "battery" and "accumulator". In this case, it makes sense to combine aspects into aspect groups.

An Aspect Sentiment Analysis of the review is a more difficult task and consists of two stages – identifying aspects and determining the sentiment of the comment on them. To complete the task of the Aspect Sentiment Analysis, we developed a simple and effective algorithm (see Figure 6). Aspects extraction based on the frequency of nouns and noun phrases mentioned in reviews based [22].

A frequency vocabulary [23] (created on text corpus) that helps to compare the obtained frequencies from reviews with frequencies from corpus is used to identify aspects. The nouns with maximum frequency deviations are claimants to be included into aspect groups. Clustering of the nouns into aspect groups was carried out manual. It should be noted,

 ISBN: 978-1-61208-358-2

that if a sentence includes nouns from several aspect groups, then it would refer to opinion about each aspect group.

The results of sentiment Analysis and Aspect Sentiment Analysis can be presented in the form of text variables $Obj = (\mathrm{Re}v_i, Sent_i, Neg_{i1}, ..., Neg_{im}, Pos_{1i}, ..., Pos_{im})$, where *Obj* – a object or a product, Rev_i – text of the *i* review, $Data_i$ – date of *i* review publication, $Sent_i$ – sentiment of *i* review, Neg_{ij} – negative sentences with mention about the *j* aspect group in the *i* review, Pos_{ij} – positive sentences with mention about the *j* aspect group in the *i* review, *i* – number of review, *j* – number of aspect group, *m* – amount of aspect groups .

Aspect extraction
Input: set of reviews *D*

1. Extract all nouns *S* from the set of reviews *D*. Count the frequency of words $\forall t = 1, |S| : f_t = N_t / N$ in the whole set of reviews D, where N – number of appearances of all words, N_t – number of appearances of the t noun.
2. Count the difference $\forall t : \Delta_t = f_t - f_t^V$ between the counted frequencies f_i and vocabulary frequencies f_t^V.
3. Sort the set of nouns *S* in descending order Δ_t.
4. Divide the set of nouns *S* from $\Delta_t > 0$ into aspect groups.

Output: set of aspect groups and aspect words

Aspect sentiment classification
Input: sentiment classifier, set of aspect groups and aspect words

1. Divide a set of reviews into set of sentences.
2. Perform sentiment classification for each sentence.
3. Check each sentence for the condition: if a sentence has a sentiment value (negative or positive) greater than a threshold *h* and contains at least one noun from any aspect group, then this sentence is labeled as an opinion (negative or positive) about the given product's aspect.

Output: positive and negative sentences with mentions about product's aspects $\{Neg_{i1}, ..., Neg_{im}, Pos_{i1}, ..., Pos_{im}\}$

Figure 5. Algorithm of Aspect Sentiment Analysis

D. Data Mining

The present paragraph suggests an algorithm of the following processing of results of sentiment analysis. The aim of the developed algorithm is to discover latent knowledge that can be used for decision support in product quality management. To realize this algorithm we use the Data Mining tool – Decision Tree, since this tool is easy to understand and interpret results. It also can explain relations between overall sentiment of review and sentiment of each aspects by means of Boolean logic.

The developed algorithm of knowledge discovery in results of sentiment analysis includes procedures presented in Figure 6. The described algorithm allows understanding of which sentiment sentences about a product's aspects influence the overall sentiment of review or, in other words, what product aspects influence customer satisfaction and in what way. The constructed Decision Tree model allows the consideration of the influence of not only separate sentiment sentences on aspects, but also their mutual presence (or absence) in the text on overall satisfaction. The Decision Tree model also allows the detection of the most significant product's aspects that are essential for the customer.

Input: positive and negative sentences with mentions about product's aspects $\{Neg_{i1}, ..., Neg_{im}, Pos_{i1}, ..., Pos_{im}\}$, vector of sentimental labels *Sent* of reviews.

1. Convert a text data $Obj = (\mathrm{Re}v_i, Sent_i, Neg_{i1}, ..., Neg_{im}, Pos_{1i}, ..., Pos_{im})$ into a boolean type by the following rules:
   2. If $Sent_i = negative$, then $newSent_i = 1$, else $newSent_i = 0$
   3. If $Neg_{ij} \neq null$, then $newNeg_{ij} = 1$, else $newNeg_{ij} = 0$
   4. If $Pos_{ij} \neq null$, then $newPos_{ij} = 1$, else $newPos_{ij} = 0$
5. Creating a decision tree where the variable $newSent_i$ is a dependent variable from $\{newNeg_{i1}, ..., newNeg_{im}, newPos_{i1}, ..., newPos_{im}\}$
6. Estimation significances of aspect groups and interpretation of extracted rules

Output: significance values of product's aspects, latent relations between satisfaction by product and satisfaction by aspects

Figure 6. Algorithm of knowledge discovery

In Figure 7 an example of Decision Tree model is presented. Nodes of the Decision Tree are the aspect variables, i.e., presence or absence in the review tonality sentences (positive or negative) with mention about some aspect from aspect group. Edges of the tree are the values of aspect variables, i.e. 1 is presence, 0 is absence. Leaves present overall sentiment of review, i.e., each branch leads to either a positive review or a negative review that meets customer satisfaction or dissatisfaction. The Decision Tree model can be expressed both in the form of Boolean functions (see Eq. 1) in a disjunctive normal form, and in natural language as a rules. Each rule is characterized by measures of reliability and support. The reliability shows what percentage of reviews containing conditions of some rule, has the same sentiment corresponding to this rule. The support shows percentage of reviews that contain conditions of some rule regarding the entire number of reviews.

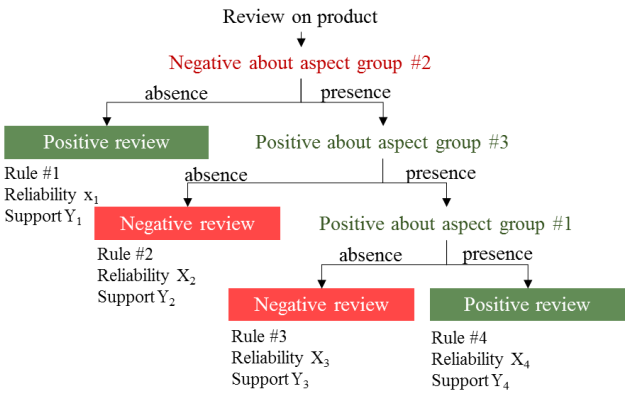


Figure 7. Example of the Decision Tree model

$$
\begin{aligned}
&Rule\ \#1:\ \overline{Neg.a.g.\#2} \rightarrow Pos.\ review\\
&Rule\ \#2:\ Neg.a.g.\#2 \cap \overline{Pos.a.g.\#3} \rightarrow Neg.\ review\\
&Rule\ \#3:\ Neg.a.g.\#2 \cap Pos.a.g.\#3 \cap \overline{Pos.a.g.\#1} \rightarrow Neg.\ review\\
&Rule\ \#4:\ Neg.a.g.\#2 \cap Pos.a.g.\#3 \cap Pos.a.g.\#1 \rightarrow Pos.\ review
\end{aligned} \quad (1)
$$

E. Measures for customer satisfaction

To measure customer satisfaction by products (or product's aspect group) we use a ratio of amount positive reviews (or positive sentences containing mentions about a product's aspect group) to all reviews (or all sentences containing mentions about product's aspect). The score of customer satisfaction *CS* by product is calculated by the formula:

$$CS = \frac{Z^{pos}}{Z^{pos} + Z^{neg}} \cdot 100\%, \quad (2)$$

where Z^{pos} – the number of positive reviews, Z^{neg} – the number of negative reviews.

The score of customer satisfaction cs_j by j product's aspect group is calculated by the formula:

$$cs_j = \frac{z_j^{pos}}{z_j^{pos} + z_j^{neg}} \cdot 100\%, \quad (3)$$

where z_j^{pos} – number of positive sentences containing mention about the j product's aspect group, z_j^{neg} – number of negative comments containing mention about the j product's aspect group.

Significance of aspects group shows how much the sentiment of a review depends on the aspect group in positive and negative sentences. Let the number of aspect groups is $g/2$, then the number of independent variables g (positive and negative statements for each group of aspect). According to the methodology described in [26] the formula for calculating the significance of variable m is:

$$Sign_m = \frac{\sum_{j=1}^{k_m}\left(E_{m,j} - \sum_{i=1}^{q_{m,j}} E_{m,j,i} \cdot \frac{Q_{m,j,i}}{Q_{m,j}}\right)}{\sum_{l=1}^{g}\sum_{j=1}^{k_l}\left(E_{l,j} - \sum_{i=1}^{q_{l,j}} E_{l,j,i} \cdot \frac{Q_{l,j,i}}{Q_{l,j}}\right)} \cdot 100\%, \quad (4)$$

where k_l – number of nodes that were split by attribute l, $E_{l,j}$ – entropy of the parent node, split by attribute l, $E_{l,j,i}$ – subsite node for j, which was split by attribute l, $Q_{l,j}$, $Q_{l,j,i}$ – number of examples in the corresponding nodes, $q_{l,j}$ – number of child nodes for j parent node.

V. EXPERIMENT

Efficacy evaluation of the developed IDSS was performed on the data obtained from 635 824 reviews about hotels in the Russian language. The reviews have been collected from the popular Internet resource tophotels.ru for the period of 2003-2013. The initial structure of the collected data consisted of the following fields: hotel name; country name; resort name; visit date; review's text; author's ratings of placement, food, and service. The data was preprocessed and loaded into the database SQL Server 2012.

Classifying sentiment used a binary scale (negative and positive) on the hypothesis that the absence of negative is positive for the product. A training set of positive and negative reviews was formed using the collected data on an author's ratings of placement, food, and service. The review site tophotels.ru uses a five-point grading scale. A review can have a maximum total rating of 15 points, and minimum total rating of 3 points. The training set included 15790 negative reviews that have 3 and 4 total points, and 15790 positive reviews that have 15 total points. We did not use the author's ratings for further data processing. Classification of another 604 244 reviews was carried out using a trained classifier.

TABLE I. COMPARISON OF METHODS FOR SENTIMENT CLASSIFICATION

#	Machine learning methods	Vector	Accuracy	
			Test No.1	*Test No.2*
1	SVM (linear kernel)	Frequency	94,2%	83,1%
2	SVM (linear kernel)	Binary	95,7%	84,1%
3	NB	Binary	96,1%	83,7%
4	NB	Frequency	97,6%	92,6%
5	NB (stop-words)	Frequency	97,7%	92,7%
6	Bagging NB	Frequency	97,6%	92,8%
7	NB (negations)	Frequency	98,1%	93,6%

For the purpose of training an effective sentiment classifier, the accuracy of classification was evaluated for machine learning methods and some peculiarities of their realization (see Table 1). The measure accuracy as a ratio of the number of correctly classified reviews to total number of reviews was used to estimate classification accuracy. Accuracy estimation was performed on two sets of data. The first set (Test No.1) represented a training set of strong positive (15 790) and strong negative reviews (15 790). Classifiers were tested by using cross validation by dividing the first set into 10 parts. The second set (Test No. 2) included random reviews from initial set of reviews (635 824) with different points (3-15 points) and was labeled manual (497 positive and 126 negative). It was used only for accuracy control of the classifier that had been trained on the first data set.

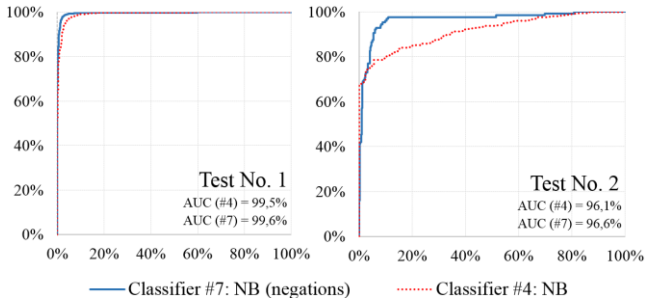


Figure 8. Comparison of ROC-curves classifiers number 7 and number 4

To estimate influence of negative particles “not” and “no”, the tagging technique was used; for example, the phrase “not good” was marked as “not_good”, and was regarded by the classifier as one word. This technique allowed the increasing of sentiment classification accuracy. Accuracy values are presented in Table 1. The most efficient ML method was naive Bayes classifier with negation techniques (#7). In Figure 8 are presented ROC-curves classifiers #4 and #7. The classifier #7 was trained on the training set and was used for further sentiment analysis.

Using the aspect extraction algorithm (Section III), we extracted the nouns that were divided into seven basic aspect groups (see Figure 9): "beach/swimming pool", "food", "entertainment", "place", "room", "service", "transport". The following step was extracting and sentiment classification sentences with words from aspect groups using classifier #7. However, not all sentences with aspects had a clearly expressed sentiment; therefore, the sentences with poorly expressed sentiment using threshold *h* were filtered out.

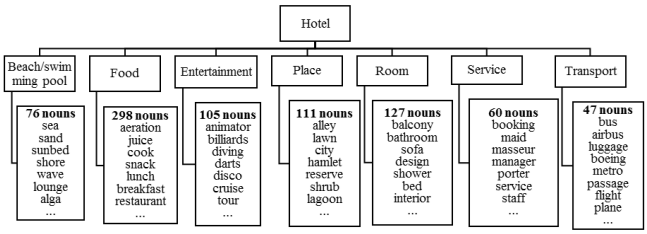


Figure 9. Aspect groups of object “hotel”

In present work we give an example of qualitative and quantitative surveys for two 5-star hotels “A” (1692 reviews) and “B” (1300 reviews) located on the resort Sharm el-Sheikh (63 472 reviews) in Egypt. Firstly, we will make a quantitative survey, measure customer satisfaction, compare it with average satisfaction in the whole resort, detect negative trends by each hotel's aspect group, and identify problems in the quality of hotels.

The dynamics of customer satisfaction is presented in Figure 10. Concerning the hotel “A”, there is a positive upward satisfaction trend from 2009, and it fixes on the average-resort level in 2013. Concerning the hotel “B”, in 2012 there was a sharp satisfaction decline and the same sharp increase in 2013. For the hotel “B”, satisfaction decrease started in June 2012, and stopped in October 2012. Then, customer satisfaction grew to the level that was higher than the average resort level being ahead of its competitor – hotel “A”.

To find reasons of hotel “B” satisfaction decrease, we will examine the diagrams in Figure 11. We can see that in 2012, the hotel “B” on average was second to the hotel “A” in such aspects as “room” (Δ12%), “place” (Δ8%), “service” (Δ5%), “beach/swimming pool” (Δ3%) and “entertainment” (Δ3%). Besides, in 2012, the Hotel “B” had more registered cases of intoxication, as well as cases of theft in August 2012. We should also note that one of the reasons of customer dissatisfaction with the hotel “B” was the initiated repair of hotel rooms and buildings which, however, paid off in 2013. Customer satisfaction with the hotel “A” aspects conforms with the average resort level.

In 2013, customer satisfaction with the hotel “B” exceeded the average level in all aspects (see Figure 12). Customer satisfaction with the hotel “A” dropped lower than average values in such aspects as “service” (Δ3%), “food” (Δ3%), “beach/swimming pool” (Δ3%) and “transport” (Δ4%). For hotel “A” manager arise questions like which aspects are the most significant for the customer and that should be improved in the first place, is it possible to “substitute” the dissatisfaction with the service, e.g., by tasty food or employ new entertainer?

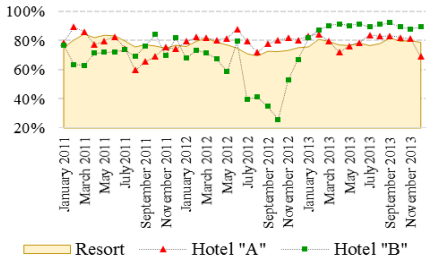


Figure 10. Dynamics of the customer satisfaction by months.

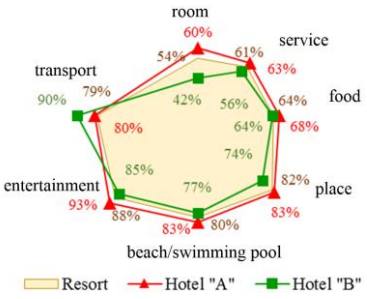


Figure 11. Comparison of the consumer satisfaction by aspects in 2012.

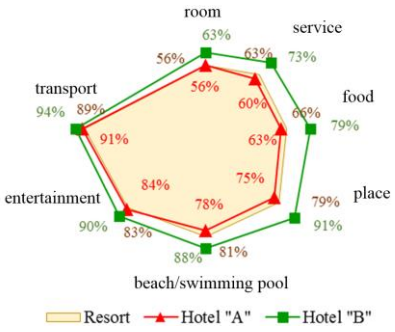


Figure 12. Comparison of the consumer satisfaction by aspects in 2013.

Decision Trees were constructed using algorithm C4.5. At the first step was constructed a tree for the all hotels of resort. Extracted rules are represented in Table 3. At the second step trees were constructed for hotel “A” and hotel “B”. Significance values of a product’s aspect groups are represented in Table 2. In Figure 13, there are the decision trees created for hotel “A” and hotel “B”. Due to the large size of the produced decision tree of the whole resort we omitted it, but in Table 3 its rules are presented that have reliability >80% and support >5% (5 rules from 27).

 ISBN: 978-1-61208-358-2

By analyzing significance values (see Table 2), we can say that the main factors of consumer dissatisfaction are a low service level (34,8%), problems with food (16%), and complaints about the hotel rooms (4%). The most critical aspect group for the hotel "B" is "room" (57,3%). In absence of negative opinions on the aspect group "room", the review would be positive with a reliability of 95,5% (see Table 3, rule #10). That is why the repair that was performed facilitated to a significant increase of consumer satisfaction. The most critical aspect group for the Hotel "A" is "service" that corresponds with the resort in a whole.

TABLE II. SIGNIFICANCE OF PRODUCT'S ASPECT GROUPS

Aspect group	Kano's model category	Sentiment of mention	Significance values		
			Resort	*Hotel "A"*	*Hotel "B"*
Service	Must-be quality	Negative	34,8%	60,2%	-
		Positive	0,7%	-	-
Food	One-dimensional quality	Negative	30,3%	27,2%	30,3%
		Positive	16%	-	-
Entertainment	Attractive quality	Negative	-	-	-
		Positive	8,5%	12,7%	12,4%
Room	One-dimensional quality	Negative	4%	-	57,3%
		Positive	2,1%	-	-
Beach/swimming pool	Attractive quality	Negative	0,2%	-	-
		Positive	2,5%	-	-
Territory	Attractive quality	Negative	-	-	-
		Positive	1%	-	-
Transport	Indifferent quality	Negative	-	-	-
		Positive	-	-	-

Using significance values, we can relate each aspect group with Kano's model categories [25]. Aspect group "service" has high significance on customer satisfaction in a "negative" case, but significance is near zero in a "positive" case (34,8% vs. 0,7%). It relates to "Must-be quality" of Kano's categories. That's why positive sentences with mentions about such aspect groups as "service" do not have an influence on sentiment of review, i.e., on overall satisfaction with hotel. That means the consumer a priori awaits a high-level service as a matter of course. For such aspect groups as "food" and "room" significance values are comparable (30,3% vs. 16% for "food" and 4% vs. 2,1% for "room"), that relates to "One-dimensional quality" of Kano's categories. Aspect groups "beach/swimming pool", "entertainment" and "territory" relates to "Attractive quality" because they have a significance on customer satisfaction in "positive" case only.

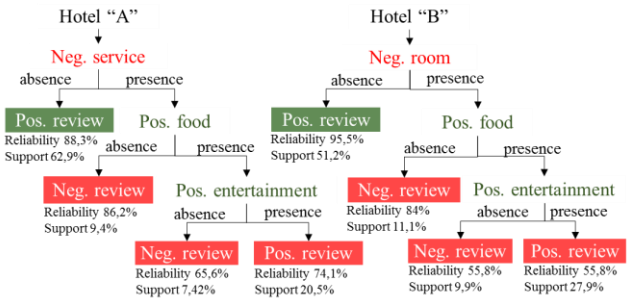


Figure 13. Decision trees for hotels

In some cases, positive mentions about "food" and "entertainment" simultaneously in a review could substitute negative mentions about "services" and provide a positive review. That's why the hotel's aspects which are contributing to customer satisfaction and important for both the resort and for the hotels are good food (30,3%) and amusing entertainment activities (8,5%). Customer satisfaction with these aspect groups can overlap dissatisfaction with "service" or "rooms" and make the customer overall satisfied (see Table 3, rules #5, #7, #11).

TABLE III. RULES EXTRACTED BY USING DECISION TREES

#	Rules	S^a	R^b
	Extracted rules on resort reviews		
1	$Food^+ \cap \overline{Service^-} \cap \overline{Food^-}$ = Positive review	37,2%	97,4%
2	$Food^+ \cap \overline{Service^-} \cap \overline{Food^-} \cap Beach^+$ = Positive review	11%	86,2%
3	$\overline{Food^+} \cap \overline{Service^-} \cap \overline{Service^-} \cap \overline{Room^-}$ = Positive review	10,6%	83,9%
4	$\overline{Food^+} \cap Service^- \cap \overline{Entertainment^+}$ = Negative review	6,9%	92,3%
5	$Food^+ \cap Service^- \cap Food^- \cap Entretainment^+$ = Positive review	5,8%	88,4%
	Extracted rules on Hotel "A" reviews		
6	$\overline{Service^-}$ = Positive review	62,9%	88,3%
7	$Food^+ \cap Service^- \cap Entertainment^+$ = Positive review	20,5%	74,1%
8	$\overline{Food^+} \cap Service$ = Negative review	9,4%	86,2%
9	$Food^+ \cap Service^- \cap \overline{Entertainment^+}$ = Negative review	7,2%	65,6%
	Extracted rules on Hotel "B" reviews		
10	$\overline{Room^-}$ = Positive review	51,2%	95,5%
11	$Food^+ \cap Room^- \cap Entertainment^+$ = Positive review	27,9%	81%
12	$\overline{Food^+} \cap Room^-$ = Negative review	11,1%	84%
13	$Food^+ \cap Room^- \cap \overline{Entertainment^+}$ = Negative review	9,9%	55,8%

a. Support. b. Reliability

The performed qualitative survey allowed the detection of the main ways to increase customer satisfaction for hotel "A". The problem aspect groups identified through quantitative survey correspond to the most significant aspects detected during the qualitative research. Hotel "A" manager should firstly increase service quality, and then increase the quality of "food" and "beach/swimming pool" maintenance. "Transport" problems – concerning flights, early check-in, and baggage storage – are not significant for customers and can be solved in the frames of service improvement. The process of service quality increase can take much time; that is why organizing entertainment and animated programs together with enhancement of restaurant service could be immediate measures for increasing customer satisfaction. Specification of managerial decisions can be performed on the basis of the information on existing problems contained in negative reviews. The extracted sentences on aspects can be directed to the appropriate hotel services.

VI. CONCLUSION AND FUTURE WORK

Poor quality of products and services contributes to a decrease of customer satisfaction. On the other hand, under the conditions of stiff competition, there are no barriers for

 ISBN: 978-1-61208-358-2

the consumer to change the supplier of goods and services. All these things can cause loss of clients and a decrease of a company's efficiency indexes. Therefore, maintaining high-quality standards should be provided by effective managerial decisions and based on opinion mining as a feedback.

The suggested conception of decision support based on the developed approach of text data processing and analysis allows performing quantitative and qualitative surveys of customer satisfaction using computer-aided procedures, and making effective managerial decisions on product quality management. The present conception allows effective reduction of labor intensity of customer satisfaction research that makes it available for use by a wide range of companies. A prototype of IDSS was developed on the basis of the suggested conception. The performed experiment has proved its efficacy for solving real problems of quality management and consistency of the results obtained. Future research on the given topic can be devoted to automatic annotating of text data, representing text amount of review in the form of a summary, and extracting useful and unique information.

REFERENCES

[1] ISO 9000-2008 The quality management system. Fundamentals and vocabulary.

[2] ISO10004:2010 Quality management. Customer satisfaction. Guidelines for monitoring and measuring.

[3] D. Gräbner, M. Zanker, G. Fliedl, and M. Fuchs, "Classification of Customer Reviews based on Sentiment Analysis", Proceedings of the International Conference in Helsingborg, Springer Vienna, Jan. 2012, pp. 460-470, ISBN 978-3-7091-1141-3.

[4] Y. Jo, A. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis", Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11), ACM New York, Feb. 2011, pp. 815-824, ISBN: 978-1-4503-0493-1.

[5] B. Lu, M. Ott, C. Cardie, and B. Tsou, "Multi-aspect Analysis with Topic Models", Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW '11), Dec. 2011, pp. 81-88, ISBN: 978-0-7695-4409-0.

[6] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web", Proceedings of the 14th international conference on World Wide Web (WWW '05), ACM New York, May 2005, pp. 342-351, ISBN: 1-59593-046-9.

[7] W. Kasper, M. Vela, "Sentiment Analysis for Hotel Reviews.", Proceedings of the Computational Linguistics-Applications Conference (CLA-2011), Oct. 2011, pp. 45–52.

[8] S. Blair-Goldensohn, K. Hannan, and R. McDonald, "Bilding a Sentiment Summarizer for Local Service Reviews", Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), March 2009, pp. 514-522.

[9] J. Ajmera, H. Ahn, M. Nagarajan, A. Verma, D. Contractor, S. Dill, and M. Denesuk, "A CRM system for Social Media", Proceedings of the 22nd international conference on World Wide Web (WWW '13), May 2013, pp. 49-58, ISBN: 978-1-4503-2035-1.

[10] H. Wang, Y. Lu, and C. Zhai, "Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10), ACM New York, July 2010, pp. 783-792, ISBN: 978-1-4503-0055-1.

[11] J. C. de Albornoz, L. Plaza, P. Gervás, and A. Díaz, "A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating", Proceedings of the 33rd European Conference on Information Retrieval (ECIR '11), April 2011, pp. 55-66, ISBN: 978-3-642-20160-8.

[12] H. Wachsmuth, M. Trenkmann, B. Stein, G. Engles, and T. Palakarska, "A Review Corpus for Argumentation Analysis", Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing, Springer (Kathmandu, Nepal), LNCS, April 2014, pp. 115-127, ISBN 978-3-642-54902-1.

[13] M. Bank, "AIM – A Social Media Monitoring System for Quality Engineering", PhD thesis, Universität Leipzig, June 2013, p. 235.

[14] J. Thomsen, E. Ernst, C. Brabrand, and M. Schwartzbach, "WebSelF: A Web Scraping Framework", Proceedings of the 12th international conference on Web Engineering (ICWE 2012), July 2012, pp. 347-361, ISBN: 978-3-642-31752-1.

[15] R. Penman, T. Baldwin, and D. Martinez, "Web scraping made simple with sitescraper", 2009. [Online]. Available from: http://sitescraper.googlecode.com.

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

[17] A. G. Pazelskaya, A. N. Soloviev, "Method of the determination emotions in the lyrics in Russian", Computer program linguistics and intellectual technologies, Issue 10 (17), 2011, pp. 510-522.

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

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

[20] N. Yussupova, D. Bogdanova, and M. Boyko, "Algorithms and software for sentiment analysis of text messages using machine learning", Vestnik USATU, T. 16-6(51), 2012, pp. 91-99.

[21] N. Yussupova, D. Bogdanova, and M. Boyko, "Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach", Proceedings of the 2nd International Conference on Advances in Information Mining and Management (IMMM2012), Venice, Italy, 2012, pp. 8-14. ISBN: 978-1-61208-227-1.

[22] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red Opal: Product-Feature Scoring from Reviews", Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 182-191. ISBN: 978-1-59593-653-0.

[23] O. N. Ljashevskaja, S. A. Sharov, "The new frequency vocabulary of Russian lexic based on the Russian National Corpus", RAS, Institut of Russian language named of V. V. Vinogradov, 2009. [Online]. Available from: http://dict.ruslang.ru/freq.php.

[24] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, "Aspect Ranking: Identifying Important Product's aspects from Online Consumer Reviews", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), June 2011, pp. 1496-1505, ISBN: 978-1-932432-87-9.

[25] N. Kano, N. Seraku, F. Takashi, and S. Tsuji, "Attractive quality and must-be quality", In The Journal of the Japanese Society for Quality Control, 1984, V. 14, pp. 39-48.

 ISBN: 978-1-61208-358-2

[26] Deductor. The Algorithm Manual (ver. 5.2.0), BaseGroup Labs, 2010. [Online]. Available from: http://www.basegroup.ru/download/guide_algorithm_5.2.0.pdf

[27] M. Taboada, J. Brooke, M. Tofiloski, K. D. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Computational Linguistics, vol. 37(2), pp. 267-307, June 2011, doi:10.1162/COLI_a_00049.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3 (4–5), pp. 993-1022, Jan. 2003, doi:10.1162/jmlr.2003.3.4-5.993.

 ISBN: 978-1-61208-358-2

WLBench: A Benchmark for WebLog Data

Ahmad Ghazal
Oracle Endeca
Oracle
El Segundo, USA
ahmad.ghazal@oracle.com

Alain Crolotte
Teradata Labs
Teradata
El Segundo, USA
alain.crolotte@teradata.com

Mohammed Al-Kateb
Teradata Labs
Teradata
El Segundo, USA
mohammed.al-kateb@teradata.com

***Abstract* — In this paper, we propose a benchmark for semi structured data based on the concept of late binding. Our proposed benchmark, called WLBench, uses weblogs as a use case. We discuss the data model, the data generation, and the queries. Furthermore, we present a proof of concept using Teradata Aster platform.**

Keywords-Benchmarking; Weblogs.

I. INTRODUCTION

Data is produced by increasing volumes of a variety of data types (i.e., structured, semi-structured, and unstructured) from sources that generate new data at a considerably high rate (e.g., click streams captured in web server logs). Data with the above described volume, variety, and velocity properties is referred to as Big Data. Big Data provides numerous new analytic and business intelligence opportunities such as fraud detection, customer profiling and churn, and customer loyalty analysis. There is a tremendous interest in Big Data from academia, industry, and a large user base. Several commercial and open source providers released a variety of products to support Big Data storage, processing, and analytics. As these products mature, there is a need to evaluate and compare their performance..

There are a few benchmarks related to Big Data, e.g., YSCB [1], CALDA [2], GridMix [3] and PigMix [4]. For the most part, these benchmarks are micro and component benchmarks. BigBench [5] is perhaps the first end-to-end Big Data benchmark. It is based on a retail store that sells products online and in stores. However, handling of semi-structured weblogs by BigBench is quite limited. In particular, its specification assumes that the weblogs contain a small number and predefined set of keys.

In this paper, we propose WLBench - a self-contained benchmark for weblogs that mandates late binding. The nature of weblogs applications makes it impossible to parse the weblogs upfront and capture their content in a structured form such as relational tables. In practice, weblogs can have hundreds or even thousands of keys from which only a small subset is used in queries. This makes it impractical to produce a schema ahead of time. Weblog queries usually involve a small number of well-defined keys which are different for each query. Hence, each query has its own schema that needs to be extracted before its execution. This concept of query-specific schema situation is called late binding. With late binding, data parsing cannot be done in advance since the schema is not known at the time data is acquired and each query has its own schema. Instead, it is carried out for each query within the query context. For example, a weblog of a retailer may have a few thousand keys and, at the same time, a query like “find the top most visited 10 products” only needs the product id information. To execute such a query, product ids need to be extracted from the weblogs and then passed over to an aggregate query that counts the number of occurrences and picks the top 10 out of those.

The contributions of the design of WLBench benchmark cover the data model, data generation, and queries. In addition, we present a proof of concept using Teradata Aster [6].

The rest of this paper is divided into the following sections. Section II presents the data model. Section III contains a detailed description of the data generation requirements. Section IV describes the queries used for the proof of concept (POC). The POC is presented in Section V. Finally, Section VI provides a conclusion together with future work directions.

II. DATA MODEL

WLBench addresses the retail business problem encountered by online vendors. Clicks are done by users of a fictitious retailer having brick and mortar as well as online stores. Users can be registered or guests and they visit the site to browse products or make purchases.

The data model is basically a set of records where each record captures a single click by a guest or a registered user. Each record consists of a set of key-value pairs that describe the corresponding click. For instance, a click by user “user1” browsing some books on 10/21/2013 at 11:30 AM is represented as follows:

userid=“user1”,productid=“books”,timestamp=“2013-10-21- 11:30”, key3=”vslue3”, key5=”value5”.

Note that key3 and key5 in the above record are generic keys to represent the keys that are not referenced by the workload but are part of the weblog records.

 ISBN: 978-1-61208-358-2

III. DATA GENERATION

We designed and developed the corresponding generation special-purpose procedure for generating weblogs. The data generator features the following key functionality. The first part of the data generation concerns users and their associate information. It produces weblogs for two groups of users - registered users and guests. Registered users sign in to the system and browse and/or buy products. The activity of a registered user is logged and associated to the user id. Guest users can browse products but are not allowed to purchase until they sign in and their activities are logged in weblog entries with no values assigned to user id. The ratio between registered users and guests can be specified to the data generator.

Generally speaking the data generation produces key value pairs. For each weblog entry, there are two sets of keys (1) fixed (or known) keys and (2) random (or unknown) keys. Fixed keys correspond to the set of attributes that are used by the query workload. The list of fixed keys is userid, itemid, webpageid, transactionid, and timestamp. The userid field identifies the user currently browsing an item itself identified by itemid on a webpage identified by webpageid. The transactionid field is assigned a value only when a user makes a purchase. The timestamp field marks the time at which the user started the current browsing or purchasing activity. The values of these keys are produced in a meaningful way. Random keys have different data types with values generated randomly and are labeled key1, key2,..., etc. We set the pool of keys to be a 100 random keys and average number of 20 keys per click. This supports the issue we highlighted before about the huge number of keys in the clicks.

The data generator has the intelligence of generating weblog entries that are amenable to forming sessions. Sessions are generated for registered users with an average number of weblog entries per session that can be specified by the user. For the proof of concept in this paper, the average number of weblog rows per non-registered user is assumed to be 4 times that of registered users. Further intelligence exists in the data generator for the distribution of values of known keys. In general, no representation is made as to what weblog size should be used.

IV. QUERIES

For our proposed benchmark, we have selected the 10 queries included below in Figure 1 below. They are expressed in English so that they can be implemented freely. The queries represent some common analytics applied to weblogs like market basket, shopping cart abandonment, session information, and affinity analysis.

- Q1: Find the 10 most browsed products.
- Q2: Find the 5 products browsed the most and not purchased.
- Q3: List users with more than 10 sessions. A session is defined as a 10-minute window of clicks by a registered user.
- Q4: Find the average number of sessions per registered user per month.
- Q5: Find the average amount of time a user spends on the retailer website.
- Q6: Find the top 10 products mostly viewed together with a given product.
- Q7: Find the 5 products mostly viewed within a month before a given product is purchased.
- Q8: For users who had products in their shopping cart but did not check out, find the average number of pages they visited during their session.
- Q9: Compare the average number of items purchased registered users from one year to the next..
- Q10: Perform affinity analysis for products purchased together.

Figure 1. List of WLBench Queries

The proposed benchmark is geared toward both Data Base Management System (DBMS) and Map Reduce (MR) [7] engines. The query set addresses the strengths of both paradigms since some of them can easily be implemented using a declarative language such as SQL (Q1, Q2 and Q4), while the other 7 queries require procedural constructs in addition to a declarative language.

As part of a standard specification rule set, the implementation will require that no initial tables be built with a priori knowledge of the fields ahead of time. Queries will need to be run on the raw data whether a table is created within the query and dropped after the query results are produced or whether the query is run on the file itself or a table with all the data fields lumped together.

V. PROOF OF CONCEPT

WLBench can be executed by traditional DBMS, MR engines like Hadoop [7], or a mix of both. There is no requirement on how click data is captured and how workload queries are executed. A system under test can choose any method to store clicks as long as no parsing is done beforehand. A DBMS may capture clicks as a table with a single column for the record text. Hadoop systems can store clicks in HDFS. Workload queries can be executed in declarative languages such as SQL [8], HQL [9] & Pig [4]. As mentioned before, some queries require procedural constructs and those can be done by User-Defined Functions (UDF) [10] or MR programs.

We executed WLBench on the Teradata Aster DBMS to illustrate the feasibility of the proposed benchmark. Clicks were captured using a simple table ClickTable where each row captures one click. The key-value text of each click is stored in a column called Payload of ClickTable. Payload is defined as a long variable character field. The workload queries were written using SQL-MR syntax which has both the declarative and procedural constructs to cover all the queries. Below is an example of a simplified SQL-MR syntax for Q3.

 ISBN: 978-1-61208-358-2

```
SELECT userid, count(*) as cnt
FROM Sessionize ( parser("userid,timestamp")
ON ClickTable)
GROUP BY userid HAVING cnt > 10;
```

Late binding is illustrated by the MR function parser which parses ClickTable and forms a table with two columns namely: userid and timestamp. The output of parser is streamed out to another MR function Sessionize which fnds the 10-minute sessions based on clicks by the same user. Finally, a count, grouped by userid is done on top of the result of the Sessionize function.

VI. CONCLUSION AND FUTURE WORK

In this paper, we laid the foundation to benchmark semi-structured data based on the late binding concept. We proposed WLBench that uses weblogs as a use case. We discussed data model, data generation, and queries and presented a proof of concept using Teradata Aster platform.

In the future, we plan on providing a full specification and a benchmark kit implemented on Aster Express, a Virtual Machine (VM) for Teradata Aster available online [6].

REFERENCES

[1] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, “Benchmarking cloud serving systems with ycsb,” in SoCC ’10. New York, NY, USA: ACM, 2010, pp. 143–154. [Online]. Available: http://doi.acm.org/10.1145/1807128.1807152

[2] 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 SIGMOD ’09. NewYork, NY, USA: ACM,2009,pp.165–178. http://doi.acm.org/10.1145/1559845.1559865

[3] GridMix.GridMixbenchmark: http://hadoop.apache.org/docs/r1.2.1/gridmix.html [retrieved: July 2014]

[4] PigMix.Benchmark: https://cwiki.apache.org/confluence/display/PIG/PigMix [retrieved: July 2014]

[5] A. Ghazal, T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H.-A. Jacobsen, “Bigbench: towards an industry standard benchmark for big data analytics,” in SIGMOD ’13. New York, NY, USA: ACM,2013,pp.1197–1208.[Online]. http://doi.acm.org/10.1145/2463676.2463712

[6] Teradata. Teradata Aster: http://www.asterdata.com/ [retrieved: July 2014]

[7] Hadoop. MapReduce: http://wiki.apache.org/hadoop/MapReduce [retrieved: july 2014]

[8] Wikipedia. SQL: http://en.wikipedia.org/wiki/SQL [retrieved: July 2014.]

 ISBN: 978-1-61208-358-2

Comparative Analysis of Data Structures for Approximate Nearest Neighbor Search

Alexander Ponomarenko, Nikita Avrelin
National Research University
Higher School of Economics
Nizhny Novgorod, Russia
Email: aponomarenko@hse.ru

Bilegsaikhan Naidan
Department of Computer and
Information Science
Norwegian University of
Science and Technology,
Trondheim, Norway
Email: bileg@idi.ntnu.no

Leonid Boytsov
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
Email: srchvrs@cs.cmu.edu

***Abstract*—Similarity searching has a vast range of applications in various fields of computer science. Many methods have been proposed for exact search, but they all suffer from the curse of dimensionality and are, thus, not applicable to high dimensional spaces. Approximate search methods are considerably more efficient in high dimensional spaces. Unfortunately, there are few theoretical results regarding the complexity of these methods and there are no comprehensive empirical evaluations, especially for non-metric spaces. To fill this gap, we present an empirical analysis of data structures for approximate nearest neighbor search in high dimensional spaces. We provide a comparison with recently published algorithms on several data sets. Our results show that small world approaches provide some of the best tradeoffs between efficiency and effectiveness in both metric and non-metric spaces.**

Keywords–nearest neighbor search; metric space; non-metric search, approximate search; small world graphs

I. INTRODUCTION

Similarity searching is a fundamental topic of computer science, which naturally appears in the different fields such as pattern recognition [1], computer vision [2], collaborative filtering [3], and so on. The goal of a similarity search is to find points from a data set that are sufficiently similar to a search pattern q, also known as a *query*.

The similarity of two data points (x and y) is computed using a distance function $d(x, y)$. The smaller the value of the distance function, the more similar (close) are the points. When d is (1) a symmetric non-negative function; (2) satisfies the triangle inequality; (3) and is equal to zero only for identical points, it is called a *metric*. If d violates any of these properties, then it is called non-metric.

In this paper, we focus on the nearest-neighbor search, where one needs to find the points whose distance from the query is the smallest among all points in the data set. A k-nearest-neighbor (k-NN) search is a generalization of the nearest-neighbors search. This generalization aims to find k points closest to the query, i.e., its k nearest neighbors. In an exact version of the problem, one is required to find all k nearest neighbors. Many exact nearest-neighbor search methods were proposed. Yet, they work well only in a low dimensional metric space. (A dimensionality of a vector space is simply a number of coordinates necessary to represent a vector: This notion can be generalized to spaces without coordinates [4]).

Experiments showed that exact methods can rarely outperform the sequential scan when dimensionality exceeds ten [5]. In a literature this problem has been dubbed as "the curse of dimensionality". Using approximate search methods, which do not guarantee retrieval of all neighbors, allows one to lift the "curse".

Non-metric spaces represent another domain where most of the proposed methods are not applicable. Compared to metric spaces, it is much harder to design exact methods for arbitrary non-metric spaces, most importantly, because the triangle inequality is violated. Whenever exact search methods for non-metric spaces do exist, they also seem to suffer from the curse of dimensionality [6][7].

Thus, the goal of approximation is two-fold: It allows us to (1) reduce the search time while obtaining reasonably accurate results; (2) answer queries for data points drawn from non-metric spaces, where properties such as the triangle inequality do not hold.

Approximate search methods can be much more efficient than exact ones, but this additional efficiency comes at the expense of a reduced search accuracy. More specifically, k points obtained by an approximate nearest-neighbor search methods might not be the k closest points to the query point. One common measure of the search accuracy is a *recall*. The recall is equal to the fraction of nearest neighbors returned by a search method.

There is a lack of evaluations that compare approximate search methods for both metric and non-metric spaces. Thus, we carry out this experimental comparison by testing several efficient benchmarks on metric and non-metric data sets. These benchmarks are compared against recently proposed method based on navigable small worlds graphs [8][9]. We measure efficiency and effectiveness for complete data sets as well as study how these characteristics depend on the number of data points.

There are several surveys covering exact nearest neighbor and range search in metric spaces, in particular, a work by Chávez et al. [4]. Many classic exact methods for metric spaces are implemented in the *Metric Spaces Library* [10]. Skopal and Bustos [11] surveyed search methods for non-metric spaces.

A *Non-Metric Space Library* is an evaluation toolkit and a similarity search library that contains efficient benchmarks for both metric (e.g., Euclidean) and non-metric spaces [12][7]. In

 ISBN: 978-1-61208-358-2

particular, the library has an approximate version of the VP-tree that was shown to be competitive [7] against the multi-probe locality sensitive hashing [13] in the Euclidean (i.e., metric) space, as well as against the bbtree [6] in the case of KL-divergence [14] and Itakura-Saito distance [15] (which are both non-metric distance functions).

The paper is organized as follows: In Section II, we describe the selected benchmarks (implemented in the *Non-Metric Space Library*); In Section III, we present evaluation results; Section IV concludes the paper.

II. Implemented Methods

A. Vantage Point Tree

The Vantage Point Tree is a hierarchical space partitioning method which uses the triangle inequality to discard partitions that cannot contain nearest neighbors [16][17]. The classic version of this method supports only an exact search in metric spaces. Yet, by stretching, i.e., relaxing, the triangle inequality [18], it is possible to support approximate nearest neighbor searching in both metric and non-metric spaces [7].

Optimal stretching coefficients were found using a simple grid search. We indexed a small database sample, executed the 10-NN search for various values of stretching coefficients and measured performance. Then, we selected coefficients resulting in the fastest search at a given recall value.

B. Permutation Methods

Permutation methods are dimensionality-reduction approaches, where each point is represented by a low-dimensional integer-valued vector called a *permutation*. To obtain the permutation, we need to select m pivots π_i (e.g., by randomly sampling data points). Then, for every point x we arrange pivots π_i in the order of increasing distance $d(\pi_i, x)$. An i-th element of the permutation vector is simply a position of the pivot i in this arrangement. For the pivot closest to the data point the value the vector element is one, while for the most distance pivot the value is m. Some of the first permutation methods were independently proposed by Chavez et al. [19], as well as by Amato and Savino [20].

A basic version of this method randomly samples pivots from the data set. Then, it computes permutations for every data point and stores permutations as an array. During the search, it scans permutations of the data points sequentially and computes a distance (usually Euclidean) between the query permutation and each retrieved permutation (representing a data point). This step generates a list of candidate data points.

Afterwards, the search method sorts candidate data points based on the distances between their permutations and the permutation of the query. A fraction of points which represent the smallest distances are compared directly against the query, using the original distance function d. The underlying idea is that while computation of the original distance can be expensive, comparing low-dimensional integer-valued vectors, i.e., permutations, is an inexpensive operation.

This basic method was improved in several ways. First of all, we need only a small fraction of the data points that have permutation closest to the query permutation. Thus, computing the complete ordering of permutations is wasteful. Instead, one can resort to incremental sorting [19]. Second, one can index permutations rather than searching them sequentially: It is possible to employ a permutation prefix tree [21], an inverted index [20], or an index designed for metric spaces, e.g., a VP-tree [22].

More recently, it was proposed to index pivot neighborhoods: For each data point, we select $numPrefix \ll m$ pivots (out of m existing pivots) that are closest to the data point. Then, we associate these $numPrefix$ closest pivots with the data point via an inverted file [23]. One can hope that for similar points two pivot neighborhoods will have a non-zero intersection.

To exploit this observation, our implementation of the pivot neighborhood indexing method retrieves all points that share at least $minTimes$ nearest neighbor pivots (using an inverted file). Then, these candidates points are compared directly against the query.

Preliminary experiments showed that, depending on a data set, one of the following permutations method was the most efficient: the basic permutation method with incremental sorting, the approximate version of VP-tree index built over a set of permutations, or a pivot neighborhood index.

C. Small World

A small world method is a variant of a navigable small world graph data structure [9]. The small world graph represents an approximation of the Delaunay triangulation [24] and its respective Voronoi partitioning [24]. In a small world graph, data points are graph nodes and edges connect close data points. Ideally, it should be possible to find nearest neighbors of any data point by following just a few graph edges.

The nearest neighbor search algorithm is, thus, a greedy search procedure that carries out several sub-searches. A sub-search starts at a random node and proceeds to expanding the set of traversed nodes by following neighboring links. The sub-search stops when we cannot find points that are closer than already found M nearest points (M is a search parameter).

Indexing is a bottom-up procedure that relies on the previously described greedy search procedure. We add points, one by one. For each data point, we find N closest points using an already constructed index. Then, we create an edge between a new graph node (representing a new point) and nodes that represent N closest points found by the greedy search. Note that the greedy search is only approximate and does not necessarily return all N nearest neighbors. Empirically, it was shown that this method often creates a navigable small world graph, where most nodes are separated by only a few edges. In that, the number of edges is typically logarithmic in the size of the data set [8].

The indexing algorithm is rather expensive and we accelerate it by running parallel searches in multiple threads. The graph updates are synchronized: If a thread needs to add edges to a node or obtain the list of node edges, it first locks a node-specific mutex. Because, different threads rarely update the same node, such synchronization creates little contention and, consequently, our parallelization approach is efficient. It is also necessary to synchronize updates for the list of graph nodes, but this operation takes little time compared to searching for N neighboring points.

D. Locality Sensitive Hashing

Locality Sensitive Hashing (LSH) is a class of methods employing hash functions that tend to have the same hash

 ISBN: 978-1-61208-358-2

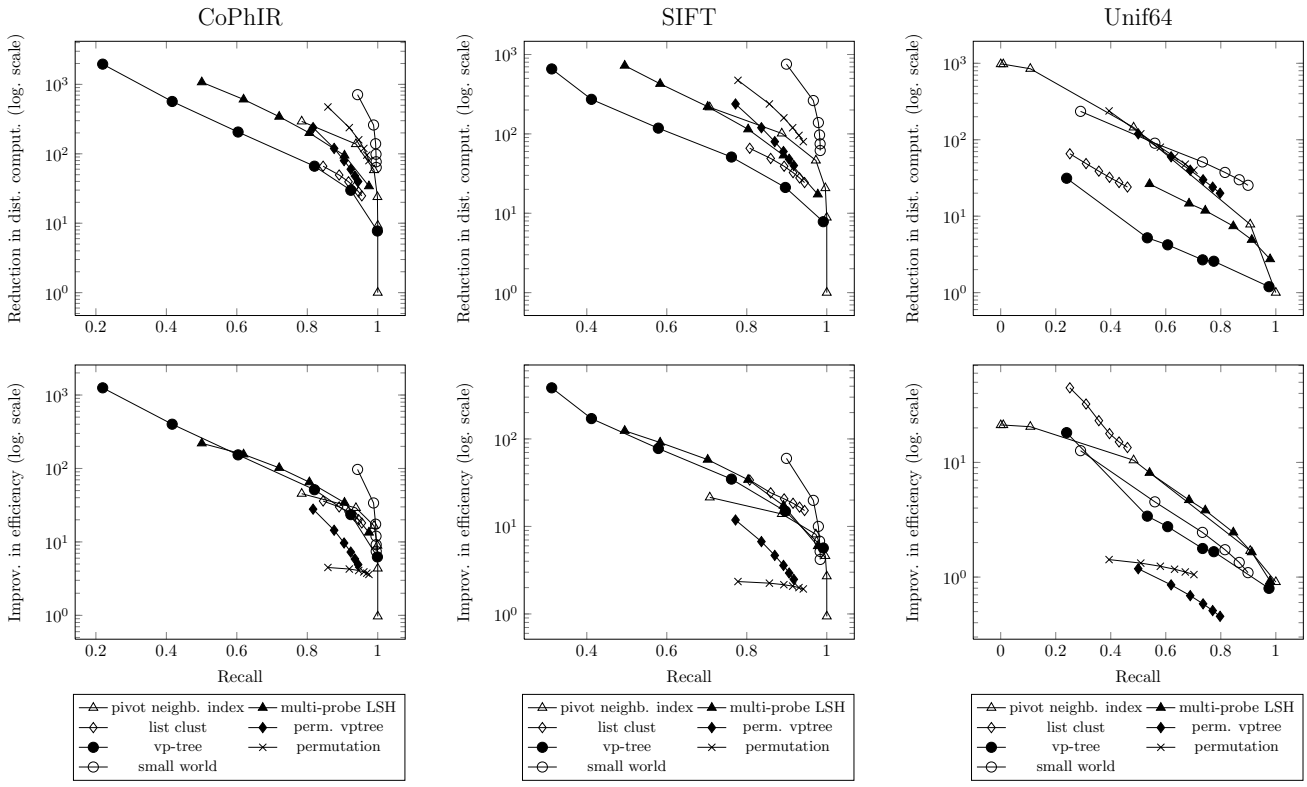


Figure 1. Performance of a 10-NN search for L_2 : plots in the same column correspond to the same data set

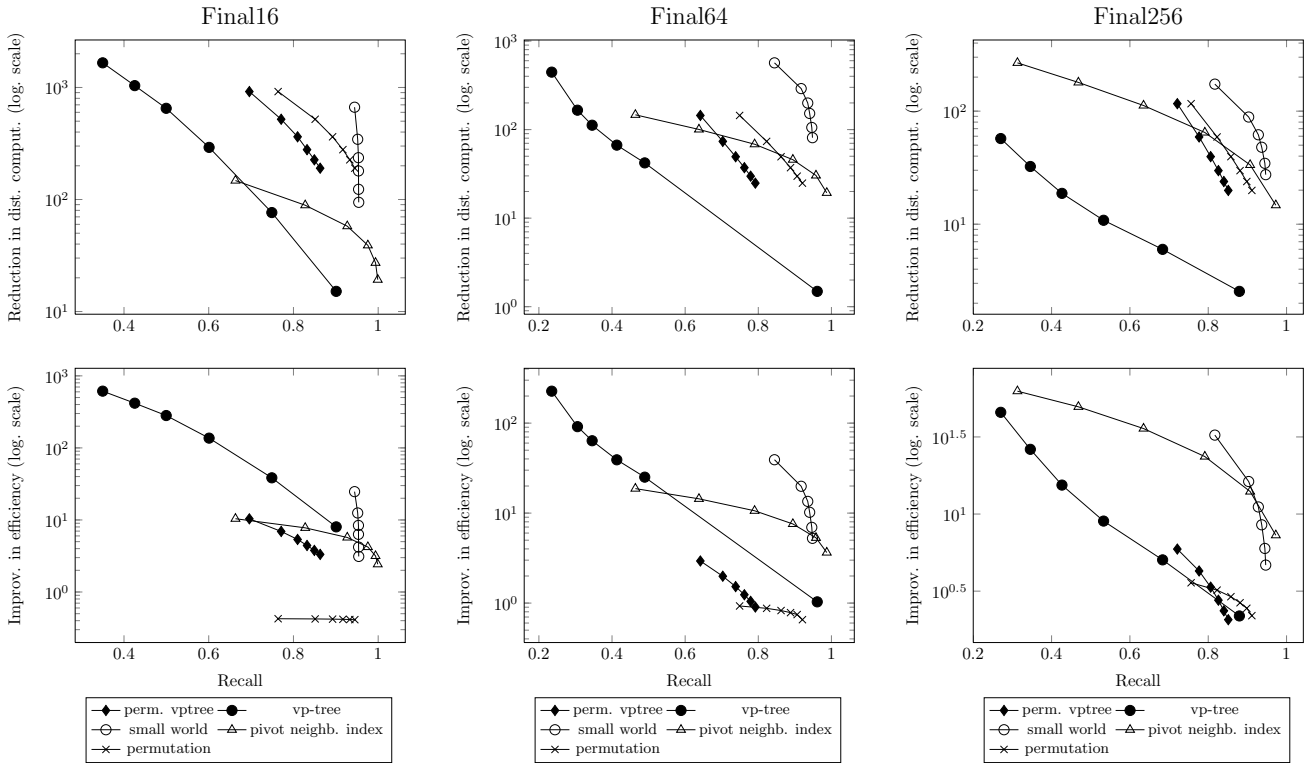


Figure 2. Performance of a 10-NN search for the KL-divergence: plots in the same column correspond to the same data set

 ISBN: 978-1-61208-358-2

values for close points and different hash values for distant points. It is a probabilistic method in which the probability of having the same hash value is a monotonically decreasing function of the distance between two points (that we compare). A hash function that possesses this property is called locality sensitive. The first LSH method was proposed by Indyk and Motwani in [25].

One drawback of this method is that it is hard to design a locality sensitive hash function for an arbitrary non-metric space. Yet, it is a very strong benchmark in the case of the Euclidean distance. This is why we used it in our experiments. More specifically, we employed a memory-efficient multi-probe LSH due to Dong et al. [13], which is implemented as a part of the LSHKIT library [13].

E. List of Clusters

The list of clusters [26] is an exact search method for metric spaces, which relies on flat (i.e., non-hierarchical) clustering. Clusters are created sequentially starting by selecting an arbitrary cluster center. Then, close points are assigned to the cluster and the clustering procedure is applied to the remaining points. Closeness is defined either in terms of the maximum distance R from the cluster center (points with distances larger than R are not included into the cluster) or in terms of the number of points N closest to the cluster center. In our work, we rely on the latter strategy and select cluster centers randomly.

The search algorithm iterates over the constructed list of clusters and checks if answers can potentially belong to the currently selected cluster (using the triangle inequality). If the cluster can contain an answer, each cluster element is compared directly against the query. Next, we use the triangle inequality to verify if answers can be outside the current cluster. If this is not possible, the search is terminated.

We modified this exact algorithm by introducing an early termination condition. The clusters are visited in the order of increasing distance from the query to a cluster center. The search process stops after vising a certain number (a method parameter) of clusters.

III. Experiments

A. Data Sets

Overall, three different distance functions were used:

- The Euclidean metric distance (L_2);
- The non-metric distance function KL-divergence [14]: $d(x,y) = \sum x_i \log \frac{x_i}{y_i}$;
- The non-metric cosine similarity: $1 - \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$.

In what follows, we summarize employed data sets and respective distance functions.

1) CoPhIR (L_2): data set is the collection of 208-dimensional vectors extracted from images in MPEG7 format [27]. Vectors are composed of five different MPEG7 features.

2) SIFT (L_2): is a part of the TexMex data set collection [28]. It has one million 128-dimensional vectors. Each vector corresponds to descriptor extracted from image data using Scale Invariant Feature Transformation (SIFT) [29].

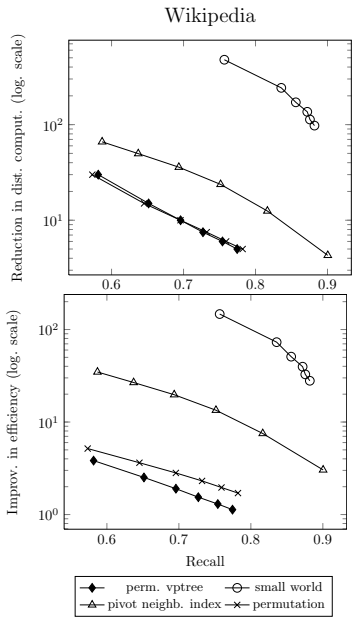


Figure 3. Performance of a 10-NN search for the 3.2 million points from the Wikipedia sparse-vector data set.

3) Wikipedia (cosine similarity): is a data set that contains 3.2 million vectors represented in a sparse format. Each vector corresponds to the TF-IDF vector of the Wikipedia page extracted using the *gensim* library [30]. This set has an extremely high dimensionality (more than 100 thousand elements). Yet, the vectors are sparse: On average only about 600 elements are non-zero.

4) Unif64 (L_2): is a synthetic data set of 64-dimensional vectors. The vectors were generated randomly, independently and uniformly in the unit hypercube.

5) Final16, Final64, and Final256 (KL-divergence): are sets of 0.5 million topic histograms generated using the Latent Dirichlet Allocation (LDA) [31]. The numeric suffix of a data set name indicates the dimensionality, which is also equal to the number of LDA topics. This data set was created by Cayton [6].

B. Evaluation

Experiments were carried out on an Linux Intel Xeon server (3.60 GHz, 32GB memory) in a single threaded mode using the *Non-Metric Space Library* as an evaluation toolkit [12]. The code was written in C++ and compiled using GNU C++ 4.7 (-Ofast optimization option).

We relied on optimized distance functions implemented with a help of SSE 4.2 SIMD instructions. In the case of the KL-divergence, further speed up are achieved by precomputing logarithms of vector elements at index time [7]. An implementation of the cosine similarity used the all-against-all comparison instruction `_mm_cmpistrm`. This implementation (inspired by the set intersection algorithm of Schlegel et al. [32]) is about 2.5 times faster than a pure C++ implementation based on the merge-sort approach.

We randomly divided a data set into two subsets. A smaller subset contained only 1000 points and was used as a query set. The remaining points were indexed. After indexing, we

 ISBN: 978-1-61208-358-2

evaluated performance of a 10-NN search (all indexes were memory-resident). To this end, a search was repeated several times to produce results at different recall values (method parameters were selected manually). This procedure was repeated five times and evaluation results were averaged over five data set splits. The variance in query times was low and, hence, we report only point estimates.

Most methods were evaluated on all data sets. Yet, the multi-probe LSH and the list of clusters were used only with the Euclidean distance. The VP-tree was not used for Wikipedia, because, due to extremely high dimensionality of this data set, the VP-tree was only marginally better than sequential searching.

Evaluation results for the Euclidean distance and for the KL-divergence are presented in Figures 1 and 2, respectively. The graphs in the first row show reduction in the number of distance computations (compared to sequential, i.e., brute force searching without an index) against the search accuracy measured by the recall (equal to the fraction of nearest neighbors returned by a method). An exact method has an ideal recall of one, which means that the exact method finds all nearest neighbors. The graphs in the second row show the overall improvement in efficiency (again, compared to sequential searching). Note that the permutation method with incremental sorting is denoted as simply *permutation* in the plots' legends.

As can be seen from the Figures 1 and 2, the small world algorithm provides the best tradeoffs between reduction in the number of distance computations and effectiveness for all three data sets. Consider, for example, the CoPhIR data in Figure 1. At the recall value of 0.9, the improvement in the number of distance computations for the small world is 1000. For all the other methods, the improvement in the number of distance computations is less than 100. To obtain a comparable improvement in the number of distance computations for, e.g., LSH, one has to tolerate the recall as low as 0.4.

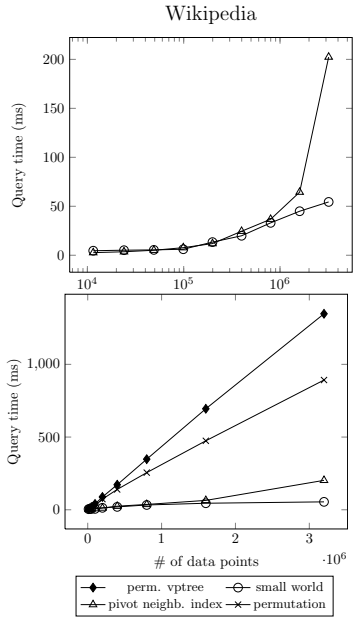


Figure 4. Dependence of the query time on the data set size. The upper plot represents only the small world method and the pivot neighborhood index. Query times are computed at roughly fixed recall values: 0.9 for the small world and 0.8 for other methods.

Typically, the larger is the reduction in the number of distance computations performed during the search, the more efficient is the method. Yet, for inexpensive distance functions (such as the Euclidean distance), the reduction in the number of distance computations does not directly translate into the overall improvement in performance. Consider the Unif64 data in Figure 1 and Final256 data in Figure 2: Despite that the small world method performs fewer distance computations than other methods in almost all the cases, the bookkeeping cost related to traversal of the small world graph can be high. As a result, the pivot neighborhood index or the multi-probe LSH are sometimes more efficient (at same recall values).

Note that both the small world and the pivot neighborhood index work well in the case of the KL-divergence (see Figure 2). For all three KL-divergence data sets, it is possible to achieve a ten-fold speed up over sequential searching while keeping the recall as high as 0.9.

The results for the complete Wikipedia data set are presented in Figure 3. The upper graph shows the reduction in the number of distance computations against the recall, while the lower graph shows the improvement in efficiency against the recall. Despite our optimized SIMD implementation of the cosine similarity is 2.5 times faster than the pure C++ version, it is still quite expensive to compute the scalar product between sparse TF-IDF vectors. As a result, in most cases, reduction in the number of distance computations maps well to the overall improvement in efficiency.

Note that the small world method is substantially better than the other methods. For example, at the recall value 0.87 it is about 40 times faster than sequential searching. The next fastest method (the pivot neighborhood index) achieves this speedup only at a significantly lower recall value of 0.6.

To measure how performance depends on the size of a data set, we also obtained results for Wikipedia subsets whose sizes varied from 12.5 thousand to 3.2 million data points. For each step and for each method we ran a search procedure several times with different options to measure performance at various values of recall (again, we manually tweaked method parameters to achieve different recall values). In the case of the small world, we selected runs that resulted in recall values closest to 0.9, while for other methods we selected runs with recall closest to 0.8. The results are presented in Figure 4, where the lower plot includes all the tested method, while the upper plot includes only the small world method and the pivot neighborhood index.

As can been seen from the Figure 4, all permutation methods have a near linear dependency for the number of distance computations on the number of data points. For the small world method, the dependence is close to being logarithmic (see the upper plot in Figure 4). In that, the small world method exhibits a greater reduction in the number of distance computations at higher recall values (0.9 vs 0.8). Compared to other permutation methods, performance of the pivot neighborhood scales much better as the number of data points increases. Yet, this method is still substantially slower and/or less accurate than the small world method.

 ISBN: 978-1-61208-358-2

IV. Conclusion and Future Work

We carried out an extensive experimental comparison using several large data sets. Our experiments involve both metric and non-metric distance functions including the challenging KL-divergence: The KL-divergence is not symmetric and does not satisfy the triangle inequality. To ease reproduction of results, we make our code publicly available, as a part of the open-source Non-Metric Space Library [12]. All data sets except CoPhIR are publicly available as well.

Our experiments show that the small world method outperforms the other methods for most recall values. Experiments with the sparse-vector Wikipedia data set demonstrate that the small world method has a near logarithmic dependence for the number of distance computation on the number of data points, which confirms previous findings [8]. That is, despite dealing with an extremely high-dimensional data set, it is possible to obtain accurate results (recall 0.9) quickly. We hypothesize that small world graph approaches are some of the most efficient high-accuracy methods in both metric and non-metric spaces.

The small world method is almost always superior in terms of the reduction in the number of distance computations. In the case of inexpensive distance functions, this does not always result in better overall performance, because traversing the small world graph can be expensive (note that the small world method is still the fastest in most cases). In the future, we plan to design a more efficient version of the small world method.

Acknowledgment

The first author is partially supported by RSF grant 14-41-00039.

References

[1] T. Cover and P. Hart, "Nearest neighbor pattern classification," Information Theory, IEEE Transactions on, vol. 13, no. 1, 1967, pp. 21–27.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic et al., "Query by image and video content: The qbic system," Computer, vol. 28, no. 9, 1995, pp. 23–32.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web. ACM, 2001, pp. 285–295.

[4] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín, "Searching in metric spaces," ACM computing surveys (CSUR), vol. 33, no. 3, 2001, pp. 273–321.

[5] R. Weber, H. J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in Proceedings of the 24th International Conference on Very Large Data Bases. Morgan Kaufmann, August 1998, pp. 194–205.

[6] L. Cayton, "Fast nearest neighbor retrieval for bregman divergences," in ICML, 2008, pp. 112–119.

[7] L. Boytsov and B. Naidan, "Learning to prune in metric and non-metric spaces," in NIPS, 2013, pp. 1574–1582.

[8] Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov, "Scalable distributed algorithm for approximate nearest neighbor search problem in high dimensional general metric spaces," in Similarity Search and Applications. Springer, 2012, pp. 132–147.

[9] ——, "Approximate nearest neighbor algorithm based on navigable small world graphs," Information Systems, vol. 45, 2014, pp. 61–68.

[10] K. Figueroa, G. Navarro, and E. Chávez, "Metric spaces library," 2007, available at http://www.sisap.org/Metric_Space_Library.html [retrieved: Jun 2014].

[11] T. Skopal and B. Bustos, "On nonmetric similarity search problems in complex domains," ACM Comput. Surv., vol. 43, no. 4, Oct. 2011, pp. 34:1–34:50.

[12] L. Boytsov and B. Naidan, "Engineering efficient and effective non-metric space library," in SISAP, 2013, pp. 280–293, available at https://github.com/searchivarius/NonMetricSpaceLib [retrieved: Jun 2014].

[13] W. Dong, Z. Wang, W. Josephson, M. Charikar, and K. Li, "Modeling lsh for performance tuning," in Proceedings of the 17th ACM conference on Information and knowledge management, ser. CIKM '08. New York, NY, USA: ACM, 2008, pp. 669–678.

[14] S. Kullback and R. A. Leibler, "On information and sufficiency," The Annals of Mathematical Statistics, vol. 22, no. 1, 03 1951, pp. 79–86. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177729694

[15] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in Proceedings of the 6th International Congress on Acoustics, vol. 17. pp. C17–C20, 1968, pp. C17–C20.

[16] J. K. Uhlmann, "Satisfying general proximity/similarity queries with metric trees," Information processing letters, vol. 40, no. 4, 1991, pp. 175–179.

[17] P. N. Yianilos, "Data structures and algorithms for nearest neighbor search in general metric spaces," in Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, ser. SODA '93. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1993, pp. 311–321.

[18] E. Chávez and G. Navarro, "Probabilistic proximity search: Fighting the curse of dimensionality in metric spaces," Information Processing Letters, vol. 85, no. 1, 2003, pp. 39–46.

[19] E. C. Gonzalez, K. Figueroa, and G. Navarro, "Effective proximity retrieval by ordering permutations," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 9, 2008, pp. 1647–1658.

[20] G. Amato and P. Savino, "Approximate similarity search in metric spaces using inverted files," in Proceedings of the 3rd international conference on Scalable information systems. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, p. 28.

[21] A. Esuli, "Use of permutation prefixes for efficient and scalable approximate similarity search," Inf. Process. Manage., vol. 48, no. 5, Sep. 2012, pp. 889–902.

[22] K. Figueroa and K. Fredriksson, "Speeding up permutation based indexing with indexing," in Proceedings of the 2009 Second International Workshop on Similarity Search and Applications, ser. SISAP '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 107–114.

[23] E. S. Tellez, E. Chávez, and G. Navarro, "Succinct nearest neighbor search," Information Systems, vol. 38, no. 7, 2013, pp. 1019–1030.

[24] H. Samet, Foundations of Multidimensional and Metric Data Structures. Morgan Kaufmann Publishers Inc., 2005.

[25] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in Proceedings of the thirtieth annual ACM symposium on Theory of computing. ACM, 1998, pp. 604–613.

[26] E. Chávez and G. Navarro, "A compact space decomposition for effective metric indexing," Pattern Recognition Letters, vol. 26, no. 9, 2005, pp. 1363–1376.

[27] P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti, "Cophir: a test collection for content-based image retrieval," arXiv preprint arXiv:0905.4627, 2009.

[28] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 1, 2011, pp. 117–128.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, 2004, pp. 91–110.

[30] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010, pp. 46–50.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, 2003, pp. 993–1022.

[32] B. Schlegel, T. Willhalm, and W. Lehner, "Fast sorted-set intersection using simd instructions." in ADMS@ VLDB, 2011, pp. 1–8.

 ISBN: 978-1-61208-358-2

A Novel Privacy Preserving Association Rule Mining using Hadoop

Kangsoo Jung, Sehwa Park, Sungyong Cho, Seog Park
Department of Computer Engineering, Sogang University, Seoul, Korea
azure84@sogang.ac.kr sehwapark@sogang.ac.kr Jsy9kr2004@hanmail.net spark@sogang.ac.kr

Abstract— **Hadoop is a popular open source distributed system that can processes large scale data. Meanwhile, data mining is one of the techniques used to find pattern and gain knowledge from data sets, as well as improve massive data processing utility when combined with the Hadoop framework. However, data mining constitutes a possible threat to privacy. Although numerous studies have been conducted to address this problem, such studies were insufficient and had several drawbacks such as privacy-data utility trade-off. In this paper, we focus on privacy preserving data mining algorithm technique, particularly the association rule mining algorithm, which is a representative data mining algorithm. We propose a novel privacy preserving association rule mining algorithm in Hadoop that prevents privacy violation without the loss of data utility. Through the experimental results, the proposed technique is validated to prevent the exposure of sensitive data without degradation of data utilization.**

Keywords-Privacy preserving data mining; Association rule mining; Hadoop.

I. INTRODUCTION

Hadoop [1] is a scalable and stable distributed data processing open-source project that has become a "de facto" standard among big data processing techniques. The development of big data processing techniques such as Hadoop has contributed to the proliferation of massive data analysis, which has not yet been explored in literature. Large size data processing utility is enhanced through its combination with data mining algorithm that is employed in rule mining and pattern recognition. Thus, numerous studies have been conducted to apply existing mining techniques to the MapReduce programming model. Meanwhile, data mining using big data may result in serious privacy violation through the inference of sensitive information. Therefore, research about privacy preserving data mining algorithm of massive datasets is necessary.

Data mining that utilizes big data requires a considerable amount of resources for proper processing. However, constructing this type of environment is burdensome for an individual or a single company. For this reason, cloud platforms, such as Amazon EC2 [2], provides service related to big data mining processes at a lower cost. However, in such platforms, personal data can flow to untrusted cloud service provider during the data mining process. The solutions proposed in literature, which include encryption [3] and privacy preserving data mining [7] algorithm; however, these methods have several disadvantages. Encryption has a too rigorous restriction because of its low computational performance, while privacy preserving data mining algorithm has one weakness on the tradeoff between privacy protection degree and data utility.

In this paper, we propose a novel Privacy Preserving Data Mining (PPDM) algorithm to overcome the limitation of existing methods with the following considerations: (1) under the environment of untrusted external cloud platform and (2) without the loss of data utility while performing privacy preserving data mining. Our method focuses on the association rule mining algorithm which is one of the data mining techniques that have received considerable attention. The proposed technique does not use encryption, but prevents the exposure of data collected under the agreement of the data provider.

The remainder of this paper is organized as follows: Section 2 reviews the association rule mining algorithm, privacy preserving data mining algorithm, and privacy preserving data processing method based on the Hadoop framework. Section 3 describes our preliminary assumption and motivation. Section 4 introduces the proposed privacy preserving association rule mining algorithm. Section 5 presents the results of the performance evaluation. Finally, Section 6 summarizes and concludes the paper.

II. RELATED WORKS

A. Association rule mining

Association rule mining [4] is a data mining algorithm that discovers interesting relations among merchandises based on purchase history. Relations are generally represented as a rule. Meanwhile, support and confidence of an itemset are used as a measure to select interesting rules from the set of all possible rules.

$$\text{Support}(X, Y) = \frac{\#\{customers\ who\ bought\ X\ and\ Y\}}{\#\{customers\}} \quad (1)$$

$$\text{Confidence}(X \rightarrow Y) = \frac{\#\{customers\ who\ bought\ X\ and\ Y\}}{\#\{customers\ who\ bought\ X\}} \quad (2)$$

The well-known apriori algorithm [5] identifies association rules by attempting to select frequent item set that has minimum support value. Frequent item sets are extended from one to the maximum length of transaction until no further extensions are found. Finding frequent itemsets, apriori algorithm determines relations by calculating the confidence value of frequent itemsets.

 ISBN: 978-1-61208-358-2

B. *Privacy preserving data mining*

Privacy preserving data mining technique can be classified into randomization and distributed privacy preserving techniques. For randomization, data perturbation and noise addition are generally used [6]-[8], and k-anonymity [8] and differential privacy [9] have been investigated recently as well. Meanwhile, distributed privacy preserving technique is a method that employs secure multi-party computation and encryption to share mining result in a distributed environment without revealing sensitive information about particular parties.

Privacy preserving mining of association rule [12][15] is an essential part in data mining. A crucial step in privacy preserving association rule is to find the global support value of frequent item set without compromising the privacy of sensitive data. To achieve this, randomization technique [13] is used, through which the data provider sends randomized data to the data miner. Data miner calculates the support value of frequent itemsets using randomized data. The support value of frequent itemset is different from the original data's support value. Hence, randomization technique prevents the exposure of correct frequent items set.

However, Evfimievski et al. [15] points out that traditional privacy preserving association rule mining has problems when it employs uniform randomization technique. Thus, they proposed a novel randomization method to overcome the existing limitation of the technique. However, a randomization technique has several disadvantages that reduce the accuracy of association rule while preserving data privacy. This issue is the privacy-data utility tradeoff problem.

C. *Massive data processing in Hadoop*

Given the growing importance of massive data processing, researchers have explored the application of data mining in a Hadoop environment. Mahout [16] is an open source library that implement machine learning algorithm in the MapReduce programming model. Mahout provides various data mining algorithm such as clustering and classification that can run without additional MapReduce programming in Hadoop.

However, the development of big data mining techniques entails the increased possibility of privacy violation. Ko et al. [17] and Zhang et al. [18] proposed a privacy preserving data processing framework based on Hadoop, and related studies have been conducted in this area. However, privacy preserving data mining algorithm in Hadoop has yet to be explored.

III. PRELIMINARIES

A. *Problem definition*

1) Assumption 1. Service providers want to exploit the external cloud service platform based on Hadoop framework to achieve association rule mining for big data.

To achieve data mining in a large dataset, using a distributed processing framework is advantageous. However, implementing a large-scale distributed environment for an individual or a single company is a rigorous task. In this regard, employing external cloud services such as Amazon EC2 is an efficient solution. For example, The New York Times uses cloud services to convert their news data into digital information. At present, most of these cloud services use Hadoop framework based on the MapReduce programming model. Hence, revising the existing data mining algorithm to correspond to the Hadoop system is necessary.

2) Assumption 2. Data providers want to limit privacy violations when the service provider processes data using external cloud services.

Service providers send data to external cloud services for data processing. However, such external cloud services are not trusted third party. Thus, if the service provider provides raw data to external cloud service without any consideration of privacy, privacy leakage of sensitive information occurs during the data mining process. As indicated in Fig. 1, although data providers agree to provide data to service providers they do not agree with offering data to an external cloud server. As such, the privacy of the data provider should be sufficiently protected.

3) Assumption 3. Service providers want to maximize the accuracy of rule which is extracted through association rule mining.

Existing methods to protect the privacy of data providers add noise to data. However, such techniques limit the performance of data mining. With the goal of service providers to protect the privacy of data providers, and to maximize the data utility, a privacy-preserving data mining algorithm, which considers the privacy-data utility trade-off, is necessary.

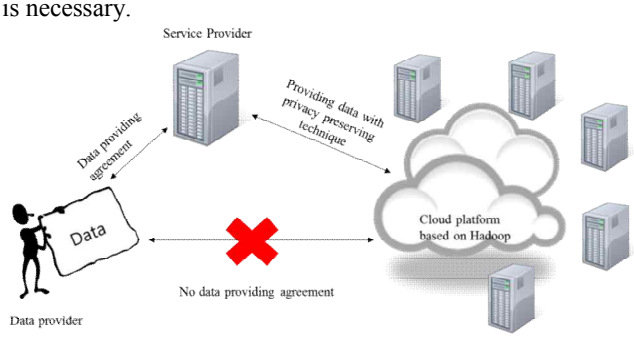


Figure 1. The proposed architecture of association rule mining in cloud platforms.

B. *Motivation*

As mentioned in Assumption 3, the randomization technique that adds noise to data results in a problem in privacy-data utility trade-off. Thus, high privacy protection degree reduces data utility, and the high data utility requirement increases the possibility of privacy violations.

In this paper, we propose a novel privacy-preserving association rule mining algorithm that prevents privacy violation in untrusted external cloud platforms without deteriorating data utility. The proposed technique adopts an identification key using prime number property to filter out

 ISBN: 978-1-61208-358-2

noise, and to process MapReduce programming for privacy-preserving association rule mining. The proposed method has several disadvantages, such as increasing calculation cost with the additional noise. Unlike privacy and data utility tradeoff, calculation cost and privacy relationship are not zero sum; Hadoop can efficiently handle the calculation cost.

IV. PRIVACY PRESERVING ASSOCIATION RULE MINING IN HADOOP

A. Basic scheme

As mentioned in Section 2, the key step of apriori algorithm is exploiting frequent itemsets. Selecting frequent itemsets results in increased calculation cost and privacy violation when performed through an external cloud service. As such, service providers add noise to the transaction data to prevent the exposure of correct frequent itemsets. However, this process results in reduced data utility. The proposed technique prevents data utility degradation by assigning an identification key that can distinguish original data from noise data. With the use of the identification key to filter out noise added by the service provider, the correct association rule can thus be extracted without utility degradation and privacy violation.

We assume that the set of transaction is $T=\{T_1,T_2,\ldots,T_n\}$, the set of item is $I=\{I_1,I_2,\ldots,I_{items}\}$, and the set of noise is $D=\{D_1,D_2,\ldots,D_{dummy}\}$. The identification key is defined as follows:

Definition 1. Identification key

We assume the set of integers $K=\{K_1,K_2,\ldots,K_n\}$, $n \geq items+dummy$. If K_i satisfies the following conditions, we define K_i as the identification key of item i.

$$key(I_i) = K_i \ (1 \leq i \leq items) \quad (3)$$

$$key(D_j) = K_{items+j} (1 \leq j) \quad (4)$$

$$K_i \neq K_j, (\forall i,j, i \neq j \ and \ K_i, K_j \in K) \quad (5)$$

The outline of the proposed method is as follows Fig. 2. The service provider adds a dummy item as noise to the original transaction data collected by the data provider. Subsequently a unique identification key is assigned to the dummy and the original items. The service provider maintains a mapping table for the item and identification key to filter out the dummy item after the selection of frequent itemset in an external cloud platform. Apriori algorithm is then performed in the external cloud platform using data sent by the service provider. The external cloud platform returns the frequent itemset results and count value to the service provider. The service provider filters the frequent itemset that is affected by the dummy item using an identification key, and extracts the correct association rule using frequent itemset without the dummy item. The extraction association rule is not a burden to the service provider, considering that the amount of calculation required for extracting the association rule is not much.

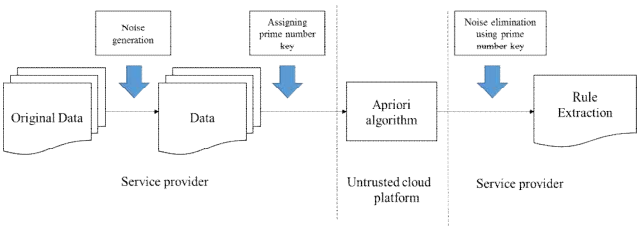


Figure 2. Overview of the proposed association rule mining

Definition 2. Prime number key

We assume the set of prime number $P=\{P_1,P_2,\ldots,P_n\}$. If P_i satisfies the following conditions, we define P_i as the prime number key of item i.

$$key(I_i) = P_i \ (1 \leq i \leq items) \quad (6)$$

$$key(D_j) = P_{items+j} (1 \leq j) \quad (7)$$

$$P_i \neq P_j, (\forall i,j, i \neq j \ and \ P_i, P_j \in P) \quad (8)$$

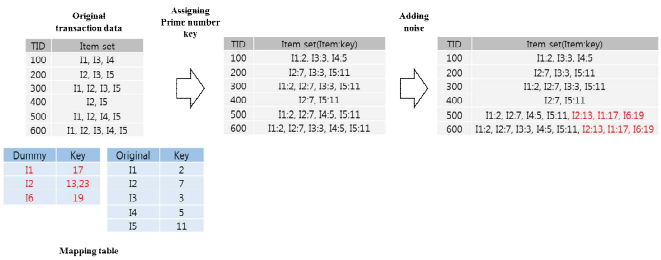


Figure 3. Prime number key assignment

B. Apriori algorithm using prime number indentification key

1) Prime number key

The method that generates dummy item with the prime number key is as shown in Fig. 3: The service provider generates dummy item as noise and assigns a prime number key to each item in the transaction data. In addition, the service provider maintains a mapping table to maintain relations with the prime number key. For example, TID500 transaction in Fig. 3 is {I1,I2,I4,I5}. The service provider assigns the prime number key to TID500, and TID500 becomes an item: prime number key pair transaction, such as {I1:2,I2:7,I4:5,I5:11}. Subsequently, a dummy item is randomly added using a noise generation algorithm, which will be explained in Section 4.3. Finally, TID500 becomes {I1:2,I2:7,I4:5,I5:11,I2:13,I1:17,I6:19}. Prime number key is used to distinguish the original item from the dummy item at the service provider, as well as a key value for MapReduce programming. The dummy item can have two or more prime number keys to enhance the degree of privacy protection.

 ISBN: 978-1-61208-358-2

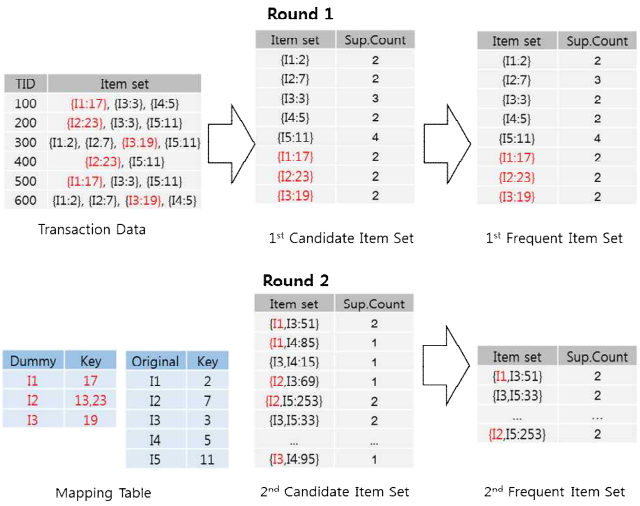


Figure 4. Apriori algorithm using prime number key

2) Apriori algorithm using a prime number key

Apriori algorithm is performed in the external cloud platform using Hadoop framework, which is similar to the traditional apriori algorithm, aside from the use of the item and prime number key pair as a key value. The frequent itemset selection process is shown in Fig. 4.

Each transaction data consists of a prime number key and an item pair. This pair and the number of pairs are regarded as the key and value, respectively, for MapReduce programming. During the processing, a subset of the transaction item's prime number key is multiplied by each item's prime number key. For example, a subset of transaction item second round frequent itemset {I4, I5} consists of I4 and I5, and the prime number key of {I4, I5} is 55, which is assigned by multiplying with I3 and I5's prime number keys, 5 and 11. Prime numbers have a property that makes multiplying with prime numbers unique. Hence, a prime number can be used as a key value for MapReduce programming. The prime number key for the subset of transaction is defined as

$$\text{key}(I') = \prod_{t=1}^{m} key(I_{r_t}) = \prod_{t=1}^{m} P_{r_t} \quad (9)$$

3) Elimination of dummy item

The frequent itemset and the number of frequent itemsets that is performed in an external cloud service platform are returned to the service provider with the prime number key. The proposed technique uses a prime number key to eliminate the frequent itemset that is influenced by the dummy item. If the frequent itemset's prime number key contains the dummy item's prime number key, the former can be divided by the latter. Hence, the service provider performs modular operation using the dummy item's prime number key to filter out frequent itemset that contains the dummy item. The filtering stage eliminates the dummy items from the remaining frequent itemsets. Using this frequent itemset, the service provider can thus extract an accurate association rule.

4) Advantage of the prime number key

The service provider can use other means to filter out noise, such as the hash value. However, this type of indexing method requires a sequential search to check the dummy item, the time complexity of which is O(n*m), (n = number of dummy items; m = number of frequent itemsets). However, the proposed technique only performs modular operation to filter out the dummy item, and time complexity is O(n), (n = number of dummy items). The disadvantage of prime number key is the size growth caused by multiplying prime numbers. We can address this weakness by using the grouping method.

C. Noise generation algorithm

In association rule mining, the possibility of privacy violation is given by the conditional probability f(X|S), which indicates the possibility of being able to infer original data X through noise data S. The proposed technique reflects the sensitivity of the item, which is set by the users to generate a dummy item as noise. That is, the data provider attaches a value from 1 to 10to the sensitivity of each item when providing transaction data. In this scale, 1 means the lowest sensitivity and 10 represents the highest sensitivity value. Item sensitivity is defined as follows:

Definition 3. Item sensitivity

We assume the set of each user's item sensitivity $S_i^m=\{S_1,S_2,\ldots,S_n\}$, and item sensitivity $\hat{S}_i$ is defined as follows:(m = # of data provider, n = # of item)

$$\hat{S}_i = \frac{(\sum_{k=1}^{m}\sum_{i=1}^{n} S_i^m)}{m} \quad (10)$$

The item sensitivity value is used to assign noise generation probability to each item. An item with a higher sensitivity has a larger possibility of generating noise, whereas an item that has lower sensitivity has a smaller possibility of generating noise. The proposed noise generation algorithm is as follows

1. The set of transaction $T_i=(I_1,I_2,..I_n)$, and integer j (j<n) is randomly selected. The probability of selecting j is proportional to the average sensitivity value of each transaction.
2. We perform Bernoulli trials to the item that is not contained in transaction T_i j times. Bernoulli trial probability ρ_i is proportional to each item sensitivity S_i.
3. The selected item is added to transaction T_i as noise.

D. Needles in haystack

Existing privacy-preserving association rule algorithms modify original transaction data through the addition of noise. However, we maintained the original transaction because our goal is to prevent data utility degradation while reducing the risk of privacy violation. Therefore, that an untrusted cloud service provider infers the real frequent itemset remains a possibility in the proposed method.

 ISBN: 978-1-61208-358-2

Despite the risk, we provide enough privacy protection because our privacy-preserving algorithm is based on "the needles in a haystack" concept. This concept is based on the idea that detecting a rare class of data, such as the needles, is hard to find in a haystack, which can be compared to a large size of data. For example, in the case of itemset {I1,I2,I3}, the possible association rules are I1→I2, I1→I3, I2→I3, I1→{I2,I3}, I2→{I1,I3}, I3→{I1,I2}, {I1,I2}→I3, {I1,I3}→I2, {I2,I3}→I1. If every possible association rule has to be extracted by adding noise, privacy protection is completely guaranteed as there is no need to extract association rule.

Existing techniques cannot add noise haphazardly because of the need to consider privacy-data utility trade-off. However, the proposed technique does not take such trade-off into consideration, for we can filter out noise using a prime number key. Nevertheless, the proposed technique incurs additional computation cost in adding noise that will make the "haystack" to hide the "needle." Therefore, we attained a trade-off between privacy and computational cost. The existing trade-off is not considered a serious problem; unlike the privacy-data utility trade-off, the computation cost and privacy protection does not have a zero-sum relationship. The problem of computational cost can be resolved by adding computing resources. Hence, the problem would be easier to be solved with the use of the Hadoop framework in a cloud environment.

V. EXPERIMENT

A. Experimental environment

In this paper, we implemented the Hadoop cluster to validate the performance of the proposed technique. The Hadoop cluster implemented in this study consists of one name node, one secondary name node, and three data nodes. The name node specifications are six-core 2.00 GHz Intex Xeon, 16 GB memory, and Ubuntu 12.10 64 bit. The specifications of the secondary name node and data node are 2-core 1.86GHz Intel CPU, 2 GB memory, and Ubuntu 12.04.3 LTS. We focused on the execution time evaluation because the proposed technique's most important drawback is the execution time degradation. We performed three experimental evaluations, namely, noise size, transaction size, and transaction length.

B. Noise size

The insertion of noise reduces system performance through the increase in amount of noise. We evaluated the performance of proposed technique by changing the noise size. We set the number of transaction as 110, where item type is 10, dummy item type is 9, and the support value is 50%. We increased noise size from 0% to 150% compared with the original transaction. The results show that the execution time increases along with the noise size. The execution time of the original transaction data where noise is not inserted is 126.008 seconds; with 50% noise, 143.578 seconds, 80%, 162.089 seconds; 120%, 172.089 seconds; and 150%, 176.040 seconds. Therefore, the execution time increases linearly in proportion to noise size. This result indicates that the proposed technique does not significantly reduce the performance.

C. Transaction size

In this experiment, we evaluated the system performance by increasing the transaction size. The item type, dummy item type, and support value were set as in the previous experiment. We increased the transaction size from 200 to 400, and set the noise size as 50% of the number of items compared with the original transaction. We can validate that in the 0% case (that is, without noise added) and 50% case (with noise added), the execution time increased because of the transaction data size. When the transaction data size is 200, the execution time of 0% is 73.586 seconds, whereas for 50%, the time is 90.587 seconds. If the transaction size is 300, 0% takes 92.6 seconds, and for 50%, the duration is 114.619 seconds. Finally, when the transaction data size is 600, the execution time of 0% is 113.598 seconds, and for 50%, 130.593 seconds.

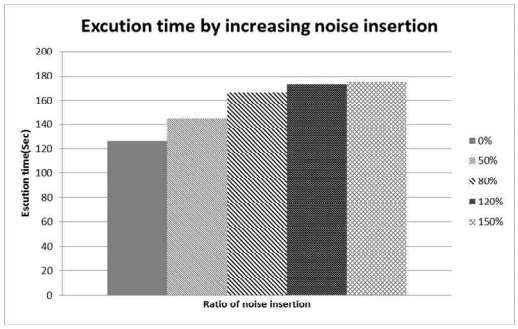


Figure 5. Execution time by increasing noise size

D. Transaction length

In this experiment, we evaluated the system performance in terms of the increasing transaction length. The transaction length is closely related to the iteration phase. If the transaction length increases, the iteration phase increases as well. The number of iteration phase directly affects the execution time. We added the dummy item to the transaction and increased the number of dummy items in the transaction to increase the transaction length. We set the item type, dummy item type, noise size, and support value as in the previous experiments. The transaction size is set as 1000.

The results show that the increase in transaction length results in the linear increase in execution time. Thus, the execution time of the proposed technique is affected by noise size, transaction size, and transaction length. However, this degradation is not a significant issue in the Hadoop framework. As such, privacy-computation cost trade-off can be achieved easily through the Hadoop framework, and the proposed technique can resolve the problem in privacy-data utility trade-off in existing randomization techniques.

 ISBN: 978-1-61208-358-2

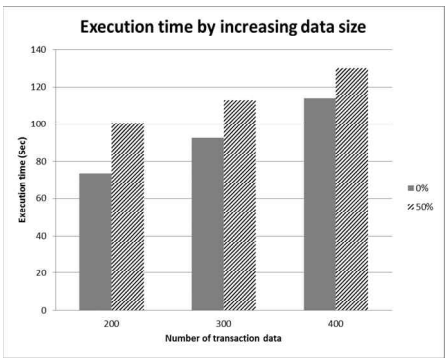


Figure 6. Execution time by data size

Figure 7. Execution time by transaction length

E. The number of Additional Rules

In this experiment, we evaluated the amount of privacy preservation of proposed technique by increasing the proportion of noise data. We increased the noise size from 15% to 120%. We set the number of transaction as 110, where item type is 10, dummy item type is 9, and the support value is 50%. We can validate that in the 0% case (that is, without noise added) and other cases (with noise added), the number of rules increased because of the ratio of noise insertion. Results of these experiments show that privacy would be protected enough.

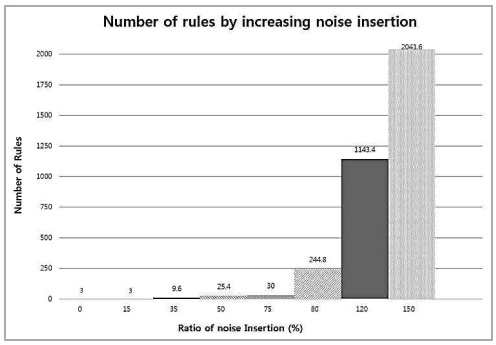


Figure 8. The Number of rules by increasing noise insertion.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel privacy-preserving association rule mining method based on the concept of "needles in a haystack." In this method, we added a dummy item as noise to the original transaction data to generate the amount of fake frequent itemset to hide the real frequent itemset during association rule mining in untrusted cloud platforms. We used a prime number key to identify the dummy and the original items, while the prime number key was used as a key value to perform MapReduce programming as well. The results showed that the proposed technique does not significantly reduce the performance while sufficiently preventing privacy violation. In future works, we will attempt to formalize the degree of privacy protection of the proposed technique and optimize the noise generation algorithm to prevent performance deterioration.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2013R1A1A2013172).

REFERENCES

[1] Apache Hadoop. http://hadoop.apache.org/.

[2] Amazon Elastic Compute Cloud. http://aws.amazon.com/ec2/.

[3] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," In Proc. of IEEE Symposium on Security and Privacy, Berkeley, CA, May. 2000, pp. 44-55

[4] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," In Proc. of Conf. Management of Data, ACM SIGMOD, Washington, DC, May. 1993, pp. 207-216.

[5] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In Proc. of 20th int. conf. on Very Large Data Bases, Santiago, Chile, Sep. 1994, pp. 487-499.

[6] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," In Proc. of Conf. on Management of Data, ACM SIGMOD, Dallas, TX, Sep. 2000, pp. 439-450.

[7] C. C. Aggarwal and P. S. Yu, "Privacy-Preserving Data Mining: A Survey," Handbook of Database Security : Application and Trends, Gertz, M. and Jajodia, S. (Eds.), 2008, pp. 431-460, Springer.

[8] K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Perturbation," In Proc. of the 5th IEEE Int'l Conf. on Data Minig, Atlanta GA, 2005, pp. 589-592.

[9] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, 2010, vol. 42, no. 4, pp. 14-53.

[10] A. Friedman and A. Schuster, "Data Mining with Differential Privacy," In Proc. of the 16th ACM Int'l Conf. on Knowledge Discovery and Data Mining, Washington, DC, Jul. 2010, pp. 493-502.

[11] A. C. Yao, "Protocols for Secure Computations," In Proc. of the 23th IEEE Symp. on Foundations of Computer Science, Chicago, Illinois, Nov. 1982, pp. 160-164.

[12] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," In Proc. of the 8th ACM Int'l Conf. on Knowledge Discovery and Data Mining, Alberta, Canada, Jul. 2002, pp. 639-644.

[13] Charu C. Aggarwal and Philip S. Yu, "A Survey of Randomization Methods for Privacy-Preserving Data Mining," Privacy-Preserving Data Mining: Models and Algorithms, Charu C. Aggarwal and Philip S. Yu, 2008, pp. 137-156, Springer.

 ISBN: 978-1-61208-358-2

[14] W. Du and M. J. Atallah, "Privacy-Preserving Cooperative Statistical Analysis," In Proc. of the 17th Conf. on Annual Computer Security Applications, New Orleans, Louisiana, Dec. 2001, pp. 102-110.

[15] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," Information Systems, 2004, VOl. 29, pp. 343-364.

[16] Apache (2013) ApacheMahout machine learning library. http://mahout.apache.org/.

[17] Ko SY, Jeon K, and Morales R, "The hybrex model for confidentiality and privacy in cloud computing," In Proceedings of the 3rd USENIX conference on hot topics in cloud computing (HotCloud'11), 2011, pp. 1-5.

[18] X Zhang, C Liu, S Nepal, C Yang, and J Chen, "Privacy Preservation over Big Data in Cloud Systems," Security, Privacy and Trust in Cloud Systems, 2014, pp. 239-257.

 ISBN: 978-1-61208-358-2