

Dhinaharan Nagamalai
Natarajan Meghanathan (Eds)

Computer Science & Information Technology

Fourth International Conference on Computer Science and Information
Technology (CoSIT 2017)
Geneva, Switzerland, March 25~26, 2017



AIRCC Publishing Corporation

Volume Editors

Dhinaharan Nagamalai,
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-921987-64-9
DOI : 10.5121/csit.2017.70401 - 10.5121/csit.2017.70418

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Fourth International Conference on Computer Science and Information Technology (CoSIT 2017) was held in Geneva, Switzerland, during March 25~26, 2017. The Fourth International Conference on Signal and Image Processing (SIGL 2017), The Fourth International Conference on Artificial Intelligence and Applications (AIAPP 2017), The Fourth International Conference on Cybernetics & Informatics (CYBI 2017), The Third International Conference on Cryptography and Information Security (CRIS 2017), The Third International Conference on Software Engineering (SEC 2017) and The Third International Conference on Data Mining and Applications (DMA 2017) was collocated with The Fourth International Conference on Computer Science and Information Technology (CoSIT 2017). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CoSIT-2017, SIGL-2017, AIAPP-2017, CYBI-2017, CRIS-2017, SEC-2017, DMA-2017 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CoSIT-2017, SIGL-2017, AIAPP-2017, CYBI-2017, CRIS-2017, SEC-2017, DMA-2017 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CoSIT-2017, SIGL-2017, AIAPP-2017, CYBI-2017, CRIS-2017, SEC-2017, DMA-2017.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai
Natarajan Meghanathan

Organization

General Chair

Natarajan Meghanathan,
Brajesh Kumar Kaushik,

Jackson State University, USA
Indian Institute of Technology - Roorkee, India

Program Committee Members

Abdulhamit Subasi	Effat University, Saudi Arabia
Ahmad Rawashdeh	University of Central Missouri, United States
Ahmed Abdou	Al-Quds University, Palestine
Ahmed Korichi	University of Ouargla, Algeria
Akash Srivastava	Indian Institute of Technology, India
Amine Laghrib	Faculté des Sciences Beni-Mellal, Morocco
Amir Rajaei	University of Velayat, Iran
Amir Rastegarnia	University of Malayer, Iran
Ankit Chaudhary	Truman State University USA
Atallah Mahmoud AL-Shatnawi	Al al-Byte University, Jordan
Azeddine Chikh	University of Tlemcen, Algeria
Chandrajit M	Maharaja Institute of Technology, India
Chin-chen Chang	Feng Chia University, Taiwan
Chuanzong Zhang	Aalborg University, Denmark
Dabin Ding	University of Central Missouri, United States
Deep Gupta	Visvesvaraya National Institute of Technology, India
Devendra Gahlot	Govt. Engineering College Bikaner, India
Elaheh Pourabbas	National Research Council, Italy
Emad Awada	Applied Science University, Jordan
Ethirajan Rajan	Pentagram Group of Companies, India
F.Kerouh,	Sherbrooke University, Algeria.
Gammoudi Aymen	University of Tunis, Tunisia
Haishan Chen	Nanfeng College of Sun Yat-Sen University, China
Hamid Alasadi	Basra University, Iraq
Hari Om	Indian Institute of Technology, India
Hayet Mouss	Batna University, Algeria
Ibtihel Nouria	Technologie and Medical Imaging Laboratory, Tunisia
Jamal El Abbadi	Mohammadia V University Rabat, Morocco
Jamel Hattay	University Of Tunis El Manar, Tunisia
Jitender Grover	Maharishi Markandeshwar University, India
John Tass	University of Patras, Greece
Jun Zhang	South China University of Technology, China
K.Santhi	Guru Nanak Institutions Technical Campus, India.
Klimis Ntalianis	Athens University of Applied Sciences, Greece
M.Keyvanpour	Alzahra University, Iran
M.V.Nageswara Rao	GMR Institute of technology, India
Mahdi Salarian	University of Illinois, USA

Manik Sharma	DAV University, India
Manish Kumar	Birla Institute of Technology and Science-Pilani, India
Manmadha Rao	GMR Institute of Technology,India
Mariusz Oszust	Rzeszow University of Technology,Poland
Masoud Vejdannik	Iran University of Science & Technology, Iran
Matineh Shaker	Northeastern University, United States
Md Zakirul Alam Bhuiyan	Fordham University, USA
Meng Ding	National Institutes of Health, USA
Mimoun Hamdi	École Nationale d'Ingénieur de Tunis (ENIT), Tunisia
Mohamedmaher Benismail	King saud University, Saudi Arabia
Mohammad alsarem	Taibah University, KSA
Mohammad Rawashdeh	University of Central Missouri, United States
Mokhtar Mohammadi	Shahrood University of Technology, Iran
Mostafa Ashry	Alexandria University, Egypt
mourchid mohammed Ibn	Tofail University Kenitra, Morocco
Naveed Ahmed	University of Sharjah,UAE
Navjot Singh	National Institute of Technology,India
Necmettin	Erbakan University, Turkey
Neda Firoz	Ewing Christian College, India
Noura Taleb	Badji Mokhtar University, Algeria
Ouafa Mah	Ouargla University, Algeria
Paulo Roberto Martins de Andrade	University of Regina, Canada
Prakash Duraisamy	University of Central Missouri, United States
Prantosh kumar Paul	Raiganj University, India
Rajlaxmi Chouhan	Indian Institute of Technology Jodhpur,India
Ramgopal Kashyap	Sagar Institute of Science and Technology, India
Ranbeer Tyagi	Maharana Pratap College of Technology,India
Razieh malekhoseini	Islamic Azad University, Iran
Sitanath Biswas	Gandhi Institute for Technology, India
Sumana	University of Rajshahi,Bangladesh
Supriya Karmakar	MellanoX Technologies,USA.
Taeghyun Kang	University of Central Missouri, United States
Tchiotso Daniel	University of Dschang,Cameroon
Upasna Vishnoi	Sr. Digital Design Engineer,USA
Vasanth Kishore	Sathyabama University, India
Vilas H Gaidhane	Birla Institute of Technology and Science, Pilani,UAE
Wonjun Lee	The University of Texas at San Antonio, USA
Xuechao Li	Auburn University, USA
Yan Lei	Beijing Forestry University,China
Yuanchang Sun	Florida International University,USA
Yueying Kao	Chinese Academy of Sciences, China
Yuriy Mishchenko	Izmir University of Economics,Turkey
Yuying Shi	North China Electric Power University,China
Zhao Peng	Huazhong University of Science and Technology, China
Zhu Jiahua	National University of Defense Technology, China.

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)



Database Management Systems Community (DBMSC)



Information Technology Management Community (ITMC)



Organized By



Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

Fourth International Conference on Computer Science and Information Technology (CoSIT 2017)

Unsupervised Detection of Violent Content in Arabic Social Media..... 01 - 07
Kareem E Abdelfatah, Gabriel Terejanu and Ayman A Alhelbawy

Investigating Binary String Encoding for Compact Representation of XML Documents..... 09 - 16
Ramez Alkhatib

Use of Adaptive Coloured Petri Network in Support of Decision Making..... 17 - 27
Haroldo Issao Guibu and João José Neto

Adaptive Automata for Grammar Based Text Compression..... 173 - 183
Newton Kiyotaka Miura and João José Neto

Fourth International Conference on Signal and Image Processing (SIGL 2017)

A Novel Adaptive - Wavelent Based Detection Algorithm for Chipless RFID System..... 29 - 38
Meriam A. Bibile and Nemai C. Karmakar

Handwritten Character Recognition Using Structural Shape Decomposition..... 39 - 47
Abdullah A. Al-Shaher and Edwin R. Hancock

Diving Performance Assessment by Means of Video Processing..... 49 - 58
Stefano Frassinelli, Alessandro Niccolai and Riccardo E. Zich

Fourth International Conference on Artificial Intelligence and Applications (AIAPP 2017)

Is AI in Jeopardy ? The Need to Under Promise and Over Deliver – The Case for Really Useful Machine Learning..... 59 - 70
Martin Ciupa

Fourth International Conference on Cybernetics & Informatics (CYBI 2017)

**Ankle Muscle Synergies for Smooth Pedal Operation Under Various
Lower-Limb Posture**..... 71 - 78
Kazuo Kiguchi, Takuto Fujita, Sho Yabunaka, Yusaku Takeda and Toshihiro Hara

**A Study on the Motion Change Under Loaded Condition Induced by
Vibration Stimulation on Biceps Brachii**..... 167 - 172
Koki Honda and Kazuo Kiguchi

Third International Conference on Cryptography and Information Security (CRIS 2017)

Advanced LSB Technique for Audio Stenography..... 79 - 86
Mohammed Salem Atoum, Mohammad M Alnabhan and Ahmad Habboush

**Securing Online Accounts via New Handshake Protocol and Granular Access
Control**..... 87 - 103
Mehrdad Nourai and Haim Levkowitz

**New Non-Coprime Conjugate Pair Binary to RNS Multi-Moduli for Residue
Number System**..... 105 - 111
Mansour Bader, Andraws Swidan and Mazin Al-hadidi

Third International Conference on Software Engineering (SEC 2017)

**Towards a Multi-Feature Enabled Approach for Optimized Expert
Seeking**..... 113 - 125
*Mariam Abdullah, Hassan Nouredine, Jawad Makki, Hussein Charara,
Hussein Hazimeh, Omar Abou Khaled and Elena Mugellini*

Estimating Handling Time of Software Defects..... 127 - 140
George Kour, Shaul Strachan and Raz Regev

Need for a Soft Dimension..... 141 - 145
Pradeep Waychal and Luiz Fernando Capretz

**Third International Conference on Data Mining and Applications
(DMA 2017)**

**The Annual Report Algorithm : Retrieval of Financial Statements and
Extraction of Textual Information..... 147 - 166**
Jörg Hering

**Storage Growing Forecast with Bacula Backup Software Catalog Data
Mining..... 185 - 196**
Heitor Faria, Rommel Carvalho and Priscila Solis

UNSUPERVISED DETECTION OF VIOLENT CONTENT IN ARABIC SOCIAL MEDIA

Kareem E Abdelfatah^{1,3}, Gabriel Terejanu¹, Ayman A Alhelbawy^{2,3}

¹Department of Computer Science and Engineering,
University of South Carolina, Columbia, SC, USA

²Computer Science and Electrical Engineering Department,
University of Essex, Essex, United Kingdom

³Computers and Information Faculty, Fayoum University, Fayoum, Egypt

ABSTRACT

A monitoring system is proposed to detect violent content in Arabic social media. This is a new and challenging task due to the presence of various Arabic dialects in the social media and the non-violent context where violent words might be used. We proposed to use a probabilistic non-linear dimensionality reduction technique called sparse Gaussian process latent variable model (SGPLVM) followed by k-means to separate violent from non-violent content. This framework does not require any labelled corpora for training. We show that violent and non-violent Arabic tweets are not separable using k-means in the original high dimensional space, however better results are achieved by clustering in low dimensional latent space of SGPLVM.

KEYWORDS

Violence, Social Media, Arabic, SGPLVM, Dimensionality Reduction, Unsupervised learning

1. INTRODUCTION

According to the Arab Social Media Report, there were 6 million Twitter users in the Arab world in March 2014, posting on average around 17 million tweets per day [1]. Twitter provides profound information as people share with others what they like and do not like, their beliefs, their political opinions, and what they observe. Due to dramatic problems plaguing much of the Arab world, a significant amount of content on social media is about violence and abuse.

Detecting offensive and violent content in social media is a very active research area, especially in the last few years. This type of research is valuable to various organizations such as Human Rights Organizations (HRO). In some crisis countries like Iraq or Syria, it may be dangerous and not safe for HROs to obtain reports and monitor the human rights situations through the usual process. Therefore, mining social media might be a solution to the problem of detecting and identifying human rights abuses safely. However, according to our knowledge there is very little work for detecting violent content in Arabic social media. This is a serious gap, as there is a real need for such kind of research in Arabic social media.

Arabic language in social media is one of the most challenges languages to be study and analyzed. Arabic is the official language in around 22 countries with more than 350 million people around the world [2]. All of these countries are Diglossia societies where both the standard form of the language, Modern Standard Arabic (MSA), and the regional dialects (DA) are used [3]. MSA is used in official settings while DA is the native tongue of Arabic speakers. DA does

not have a standard orthography and it is divided into several groups among these countries [4]. Nowadays, these dialects are extensively utilized in social media text, in spite of their original absence from a written form [3].

Detecting violence content in Arabic social media is not a trivial task. Not only because the different Arabic dialects that we have mentioned above, but also because of violent Arabic words are not always representative of violent context. For example, the word “Killing” has both a violent meaning but it may also be used in a non-violent context as in the following tweet examples [5].

إن الذاكرة والألم توأمان لا تستطيع قتل الألم دون سحق الذاكرة

"The memory and the pain twins, you cannot kill the pain without crushing the memory"

تستطيع قتل الأزهار ولكن لا تستطيع أن تمنع قدوم الربيع

"You may kill the flowers but cannot prevent the arrival of spring"

On other hand, the same word can be used in a violent context, like the following example [5]:

مقتل خمسة أشخاص برصاص مسلحين والقبض على ستة مشتبه بهم

"The killing of five people shot dead by gunmen and arrested six suspects"

In this work, we tackle this problem using a recently released dataset that contains 16234 manually annotated Arabic tweets [5]. It contains different violent context like killing, raping, kidnapping, terrorism, invasion, explosion, or execution, etc. According to our knowledge this is the first study conducted on this dataset. We use an unsupervised technique to binary cluster this dataset to violent and non-violent content. First, the Sparse Gaussian Process Latent Variable Model (SG- PLVM) [6] is used as an unsupervised probabilistic non-linear Dimensionality Reduction (DR) model. Then we apply k-means on the features extracted in the previous step. Using recent released Arabic dataset [5], our experiments show that violent and non-violent Arabic tweets are not separable using k-means in the original high dimensional space, however better results are achieved using low dimensional projections provided by the SGPLVM.

2. PREVIOUS WORK

There is much research work in detecting violent content on web [7, 8]. Computer vision techniques have been proposed to detect violence in videos [9–11]. On the other hand, text mining techniques have been used to detect violence in English social media; but little work targets this problem in Arabic social media.

A probabilistic violence detection model (VDM) is proposed in Ref. [12] to extract violence related topics from social media data. The authors propose a weakly supervised technique and they used OpenCalais with Wikipedia documents, and Wikipedia and YAGO categories to build a training corpus. The dataset was built to detect violence categories such as Crimes, Accidents, War Conflict, etc. Non-violence related categories are anything other than violence, like Education and Sports. We tested OpenCalais, but unfortunately it does not support Arabic text. Also, the number of documents under violence categories in Arabic Wikipedia is very small.

Lexical Syntactical Feature (LSF) [13] has been introduced to detect offensive language in social media. The proposed system uses the user profile to get some information about the user's

English writing style. A set of lexical features like Bag of Words and N-grams, and hand-authoring syntactic rules are used to identify name-calling harassments. In additions, a users potentiality to send out offensive content in social media has been predicted using some features like style, structure, and context-specific features. This proposed method uses Naive Bayes and SVM techniques to train a classifier.

3. CLUSTERING IN A LOWER SPACE

It is very common in NLP to have a really high dimensional feature vectors. Using unsupervised techniques for clustering patterns is good and cheap choice. k-means algorithm is one of the good candidates for unsupervised learning techniques. But, k-means can give better results when it is applied on low dimensional features [14] Therefore, it is common to project a high dimensional data set onto a lower dimensional subspace using unsupervised DR techniques such as Principle Components Analysis (PCA) [15] to improve learning. It is widely used approach to project data onto a lower dimensional subspace using PCA then use k-means to cluster the data in the lower dimensions space [15].

Because unsupervised clustering algorithms such as k-means operate mainly on distances, it is vital to use a DR technique that is able to preserve the distance metric between the data points in the low dimensional subspace. PCA is the most widely used linear DR for obtaining a lower dimensional representation of a data set. PCA may maintain the dissimilarity [14] which can help the K-means to achieve better separation for clustering. We meant by preserve the dissimilarity is the ability to preserve the points that are far apart in data space to be far apart in the latent space. However, due to linearity, PCA may not capture the structure of the data through a low dimensional embedding [16].

Gaussian process latent variable model (GPLVM) [17] is a flexible non-linear approach to probabilistic modelling data in high dimensional spaces. It can be used as DR method which maps between the observed data points $Y \in \mathfrak{R}^{N \times D}$ and latent unobserved data points $X \in \mathfrak{R}^{N \times q}$. One of its advantages it can preserve the dissimilarity and smoothness between the data in high and low dimension spaces. Smoothness means that if two points in the latent space are close (far) to each other then they will be mapped to two points that are relatively close (far) to each other in the data space. The GPLVM as a probabilistic approach models the relationship between latent variables and the observed data through non-linear parametrized function $y_{:,i} = f(X, w_{:,i}) + \epsilon_{:,i}$ where $y_{:,i} \in \mathfrak{R}^{N \times 1}$ represents one dimension of the observed data and $w_{:,i} \in \mathfrak{R}^{1 \times D}$ is one row of the parameters $W \in \mathfrak{R}^{q \times D}$ which it has a prior Gaussian distribution over each of its row with zero mean and unit variance $w_i \sim N(w_i | 0, I)$ and noise $\epsilon_{:,i} \sim N(0, \sigma^2 I)$. GPLVM assumes that there is independency across the data dimensionality. Thus, the likelihood for all dimensions can be written as a product of the likelihood of the D observed dimensions.

$$p(Y | X) = \prod_{i=1}^D N(y_{:,i} | 0, K + \sigma^2 I)$$

Inferencing the latent projects can be achieved through maximizing the marginal log-likelihood of the data,

$$\log p(Y|X) = \frac{-D}{2} \log |K| - \frac{1}{2} \text{Tr}(K^{-1}YY^T) + C$$

Here, C is a constant and $K \in \mathfrak{R}^{N \times N}$ is a kernel matrix that is calculated from the training data. There are different kernel functions available that can be used. In our experiments we used the radial basis function (RBF),

$$k(x_i, x_j) = \theta_{\text{rbf}} \exp\left(\frac{-(x_i - x_j)^T(x_i - x_j)}{2\gamma^2}\right)$$

where θ_{rbf} , γ are the parameters of the kernel.

However, a major drawback with the standard GPLVM approach is its computational time. To infer the latent variable X , GPLVM uses a gradient based iterative optimization of the log likelihood which requires $O(N^3)$ complexity due to the inverse of K [6]. Therefore, the Sparse-GPLVM (SGPLVM) [6] comes to solve this issue by reducing the complexity to $O(u^2N)$ where u is the number of points retained in the sparse representation. Therefore, using Sparse-GPLVM before K-means can guarantee to preserve the dissimilarity between the data points in the latent space which leads to coherent patterns that can be detected easily via clustering.

4. DATASET

A manually annotated dataset of 16,234 tweets are used for training and testing [5]. Every tweet had been classified by at least five different annotators. As every tweet is classified by different users, it may be assigned different classes. So, a final aggregate class is assigned based on a class confidence score as it is described in the original publication [5]. In our experiments we have kept only the tweets have a confidence score more than 0.7.

Table 1: Dataset Details

Class	Training	Testing	Total	%
Violence	5673	2759	9332	57.5
Non-Violence	4790	2112	6902	42.5
Total	11363	4871	16234	

The original dataset is classified into seven violence classes: crime, violence, human rights abuse, political opinion, crisis, accidents, and conflict. There is an additional class “other”, which contains non-violence tweets where some violence words had been mentioned.

Because we are interested in detecting the violence acts in Arabic social media regardless the type of violence, all violence classes are mapped to one class “violence”, while the “other” class is mapped to “non-Violence” class. Around 70% of the dataset is used for training and 30% is used for testing as shown in Table 1.

5. EXPERIMENTS SETUP

The Arabic has a complex morphological structure especially for Dialectal Arabic [18]. Until now, there are no available standard resources for Arabic text mining and morphological analysis [18]. However for our study, we use MADIMARA [18] analysis tool because it has most of

common tools for morphological analysis in Arabic. After removing Arabic stop words and web links, we used MADIMARA to extract some morphological features like gloss and token.

Tweets are represented in a Vector Space Model (VSM) with TF-IDF weighting scheme. The baseline approach is to cluster the dataset in the original high dimensional space into two clusters using k-means [19] with different features. Then, two different DR techniques (PCA and SGPLVM) are applied. We study the ability to separate these data points in the latent space by clustering the data into two clusters using k-means. We have tried to reduce the data to different dimension (Dim) spaces and reported some of these results. Two sets of experiments have been carried out. The first set is using the Gloss feature where the original space is 14,621 dimensions. So we reduced it to 11,000, and 8000 with PCA.

Another experiment has been carried out with reducing dimensionality to 8000 but using GPLVM. The second set has been carried out using the token feature where the dimensionality is much higher i.e. 44,163 features. PCA had been used to reduce it to 35,000 and 8,000 features. For comparability reasons, GPLVM also used to reduce the dimensionality to 8000 again. To measure the performance for the clustering steps whether in the data space or latent space, we used the training data to assign each cluster to one class which maximizes the precision of “violence” class. Then, we use the precision (P), recall (R), and F-score (F) as an evaluation metric to compare between these different techniques.

6. RESULTS

Table 2 shows the results for applying k-means on the original data space and after reducing the dimensionality using PCA and SGPLVM with different features.

Gloss as a linguistic feature has a level of abstraction which multiple tokens may have the same gloss. It is noiseless as comparable to token feature each new token is considered as a new dimension.

From the gloss results, when we try to reduce the dimension using PCA to different levels (reduced to more than 45% of the original dimension), k-means is still able to sustain its performance. In two cases (higher and lower dimension), k-mean can achieve precision around 47%. However, in the token case, we can see that PCA is affected by the data representation which is noisy.

Table 2: Experimental results.

Gloss Feature				
Model	Dim	P	R	F
K-means (in data space)	14,621	0.46	0.65	0.54
PCA + K-means	11,000	0.47	0.66	0.55
PCA + K-means	8000	0.46	0.63	0.54
SGPLVMx + K-means	8000	0.56	0.60	0.58
Token Feature				
Model	Dim	P	R	F
K-means (in data space)	44,163	0.50	0.75	0.60
PCA + K-means	35,000	0.56	0.98	0.71
PCA + K-means	8000	0.49	0.72	0.58
SGPLVMx + K-means	8000	0.58	0.55	0.56

On the other hand, according to the precision metric, SGPLVM can help k-means to achieve better results than both of what it can obtain in the original space and in the lower space using PCA. Using SGPLVM for gloss features, we can increase the precision accuracy around 21% comparable to what we can get from other methods. However using token features, we tremendously decreased the dimensionality using SGPLVM to study if it is able to keep the distance between the points. Unlike the PCA, the results show that the non-linearity of SGPLVM with k-means is still able to outperform the k-means in the high data space.

7. CONCLUSION

In this paper we tackled a new challenging problem in Arabic social media. We introduced an unsupervised framework for detecting violence in the Arabic Twitter. We use a probabilistic non-linear DR technique and an unsupervised cluster algorithm to identify violent tweets in Arabic social media dataset. We compare k-means as a baseline with the results of SGPLVM and PCA with k-means. The preliminary results show that detecting violent content in this dataset using unsupervised techniques can be achieved using a lower dimensional representation of the data with results better than applying clustering on the original data set. More experiments will be carried out to achieve better results.

REFERENCES

- [1] R. Mourtada and F. Salem, "Citizen engagement and public services in the arab world: The potential of social media," Arab Social Media Report series,, 2014.
- [2] F. Sadat, F. Kazemi, and A. Farzindar, "Automatic identification of arabic language varieties and dialects in social media," Proceedings of SocialNLP, 2014.
- [3] H. Elfardy and M. T. Diab, "Sentence level dialect identification in arabic.," in ACL (2), pp. 456–461, 2013.
- [4] N. Y. Habash, "Introduction to arabic natural language processing," Synthesis Lectures on Human Language Technologies, vol. 3, no. 1, pp. 1–187, 2010.
- [5] A. Alhelbawy, P. Massimo, and U. Kruschwitz, "Towards a corpus of violence acts in arabic social media," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), (Paris, France), European Language Resources Association (ELRA), 2016.
- [6] N. D. Lawrence, "Learning for larger datasets with the gaussian process latent variable model," in International Conference on Artificial Intelligence and Statistics, pp. 243–250, 2007.
- [7] E. Whittaker and R. M. Kowalski, "Cyberbullying via social media," Journal of School Violence, vol. 14, no. 1, pp. 11–29, 2015.
- [8] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime, "TextMining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.
- [9] E. B. Nieves, O. D. Suarez, G. B. Garc'ia, and R. Sukthankar, "Violence detection in video using computer vision techniques," in Computer Analysis of Images and Patterns, pp. 332–339, Springer, 2011.
- [10] F. D. de Souza, G. C. Ch'avez, E. A. do Valle, and A. de A Araujo, "Violence detection in video using spatiotemporal features," in Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on, pp. 224–230, IEEE, 2010.

- [11] A. Datta, M. Shah, and N. D. V. Lobo, "Person-on-person violence detection in video data," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 1, pp. 433–438, IEEE, 2002.
- [12] A. E. Cano Basave, Y. He, K. Liu, and J. Zhao, "A weakly supervised bayesian model for violence detection in social media," in *Proceedings of International Joint Conference on Natural Language Processing*, pp. 109–117, Asian Federation of Natural Language Processing, 2013.
- [13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pp. 71–80, IEEE, 2012.
- [14] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, p. 29, ACM, 2004.
- [15] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in *Advances in neural information processing systems*, pp. 1057–1064, 2001.
- [16] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *The Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.
- [17] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," *Advances in neural information processing systems*, vol. 16, no. 3, pp. 329–336, 2004.
- [18] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological analysis and disambiguation for dialectal arabic.," in *HLT-NAACL*, pp. 426–432, 2013.
- [19] J. A. Hartigan and M. A. Wong, "Algorithm 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

INTENTIONAL BLANK

INVESTIGATING BINARY STRING ENCODING FOR COMPACT REPRESENTATION OF XML DOCUMENTS

Ramez Alkhatib¹

Department of Computer Technology, Hama University, Hama, Syria

ABSTRACT

Since Extensible Markup Language abbreviated as XML, became an official World Wide Web Consortium recommendation in 1998, XML has emerged as the predominant mechanism for data storage and exchange, in particular over the World Web. Due to the flexibility and the easy use of XML, it is nowadays widely used in a vast number of application areas and new information is increasingly being encoded as XML documents. Because of the widespread use of XML and the large amounts of data that are represented in XML, it is therefore important to provide a repository for XML documents, which supports efficient management and storage of XML data. Since the logical structure of an XML document is an ordered tree consisting of tree nodes, establishing a relationship between nodes is essential for processing the structural part of the queries. Therefore, tree navigation is essential to answer XML queries. For this purpose, many proposals have been made, the most common ones are node labeling schemes. On the other hand, XML repeatedly uses tags to describe the data itself. This self-describing nature of XML makes it verbose with the result that the storage requirements of XML are often expanded and can be excessive. In addition, the increased size leads to increased costs for data manipulation. Therefore, it also seems natural to use compression techniques to increase the efficiency of storing and querying XML data. In our previous works, we aimed at combining the advantages of both areas (labeling and compaction technologies), Specially, we took advantage of XML structural peculiarities for attempting to reduce storage space requirements and to improve the efficiency of XML query processing using labeling schemes. In this paper, we continue our investigations on variations of binary string encoding forms to decrease the label size. Also We report the experimental results to examine the impact of binary string encoding on reducing the storage size needed to store the compacted XML documents.

KEYWORDS

XML Compaction, XML Labeling, XML Storage, Binary encoding

1. INTRODUCTION

The ability to efficiently manage XML data is essential because the potential benefits of using XML as a representation method for any kind of data. There have been many proposals to manage XML documents. However, XML Labeling and compaction techniques are considered as two major approaches able to provide robust XML document storage and manipulation.

¹ Part of this work was done while the author was member of the Database and Information Systems Research Group, University of Konstanz

Since the logical structure of an XML document is an ordered tree consisting of tree nodes that represent elements, attributes and text data, establishing a relationship between nodes is essential for processing the structural part of the queries. Therefore, tree navigation is essential to answer XML queries. However standard tree navigations (such as depth- first or breadth-first traversals) are not sufficient for efficient evaluation of XML queries, especially the evaluation of ancestor and descendant axes. For this purpose, many node labeling schemes have been made. The use of labeling schemes to encode XML nodes is a common and most beneficial technique to accelerate the processing of XML queries and in general to facilitate XML processing when XML data is stored in databases [15].

The power of XML comes from the fact that it provides self-describing capabilities. XML repeatedly uses tags to describe the data itself. At the same time this self-describing nature of XML makes it verbose with the result that the storage requirements of XML are often expanded and can be excessive. In addition, the increased size leads to increased costs for data manipulation. The inherent verbosity of XML causes doubts about its efficiency as a standard data format for data exchange over the internet. Therefore, compression of XML documents has become an increasingly important research issue and it also seems natural to use compression techniques to increase the efficiency of storing and querying XML data [3, 4, 6, 8]. In our works, we focused on combining the strengths of both labeling and compaction technologies and bridging the gap between them to exploit their benefits and avoid their drawbacks to produce a level of performance that is better than using labeling and compression independently.

In this paper, we continue our investigations on variations of binary encoding forms that would provide for opportunities to further minimize the storage costs of the labels. The rest of the paper is structured as follows: Section 2 and 3 review The CXQU and CXDLS compaction approaches respectively. In Section 4, we present variations of binary encoding schemes can be used to minimize the storage costs of the labels. Experimental results to study the impact of prefix free encoding schemes on reducing the storage size are presented in Section 5. Finally, we conclude and outline future work in Section 6.

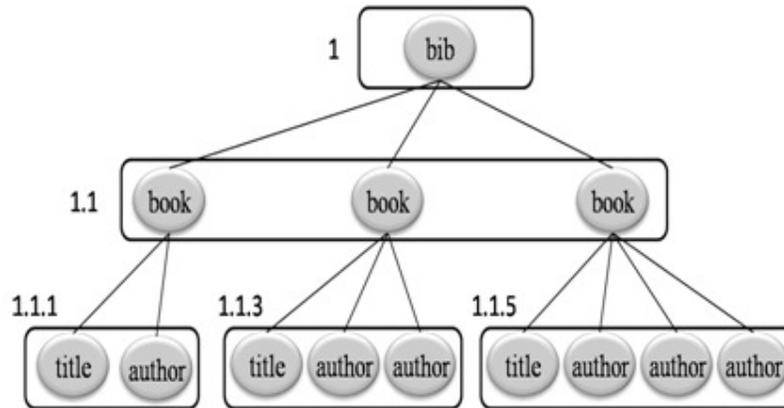


Figure 1. Simple XML document with cluster labels

2. THE CXQU COMPACTION APPROACH

CXQU is our proposed approach [1] to represent XML documents. It not only supports queries and updates but also compacts the structure of an XML document based on the exploitation of repetitive consecutive tags in the structure of the XML documents by using our proposed labeling scheme called Cluster Labeling Scheme (CLS) [1]. CLS assigns a unique identifier to each group

of elements which have the same parent (i.e. sibling element nodes). CLS preserves the hierarchal structure of XML documents after the compaction and supports the managing compacted XML documents efficiently. It allows insertion of nodes anywhere in the XML tree without the need for the subsequent relabeling of existing nodes. To compact an XML document with CXQU, first, it separates its structural information from the content to improve query processing performance by avoiding scans of irrelevant data values. CXQU then compacts the structure using our algorithm, which basically exploits the repetition of similar sibling nodes of XML structure, where “similar” means: elements with the same tag name. CXQU stores the compacted XML structure and the data separately in a robust compact storage that includes a set of access support structures to guarantee fast query performance and efficient Updates. Figure 1 displays the cluster labels and Figure 2 displays the compacted structure of a simple XML document, where the crossed-out nodes will not be stored.

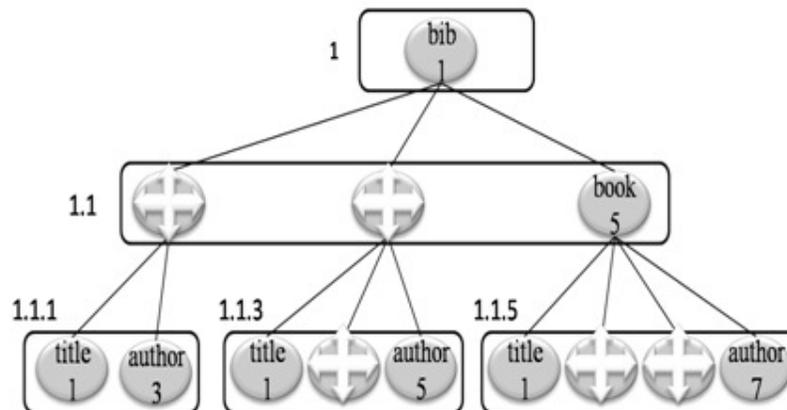


Figure 2. The compacted structure using CXQU

3. THE CXDLS COMPACTION APPROACH

We also proposed an improved technique called CXDLS [2] combining the strengths of both labeling and compaction techniques. CXDLS bridges the gaps between numbering schemes and compaction technology to provide a solution for the management of XML documents that produces better performance than using labeling and compaction independently. CXDLS compacts the regular structure of XML efficiently. At the same time, it works well when applied to less regular or irregular structures. While this technique has the potential for compact storage, it also supports efficient querying and update processing of the compacted XML documents by taking advantage of the ORDPATH labeling scheme. ORDPATH [14] is a particular variant of a hierarchical labeling scheme, which is used in Microsoft SQL Server's XML support. It aims to enable efficient insertion at any position of an XML tree, and also supports extremely high performance query plans for native XML queries.

CXDLS helps to remove the redundant, duplicate subtrees and tags in an XML document. It takes advantage of the principle of separately compacting structure from data and it also uses the ORDPATH labeling scheme for improving the query and update processing performance on compacted XML structures.

In CXDLS, the XML structure is compacted based on the basic principle of exploiting the repetitions of similar nodes in the XML structure, where two nodes N and N' of XML structure are said to be „similar“ if they are consecutive elements, i.e. sibling nodes, in the structure and have exactly the same tag name. Another principle is to exploit the repetitions of identical

subtrees, where two subtrees S and S' of XML structure are said to be „identical“ if they are consecutive and have exactly the same structure. Figure 3 shows the ORDPATH labels and Figure 4 displays the compacted structure using CXDLS.

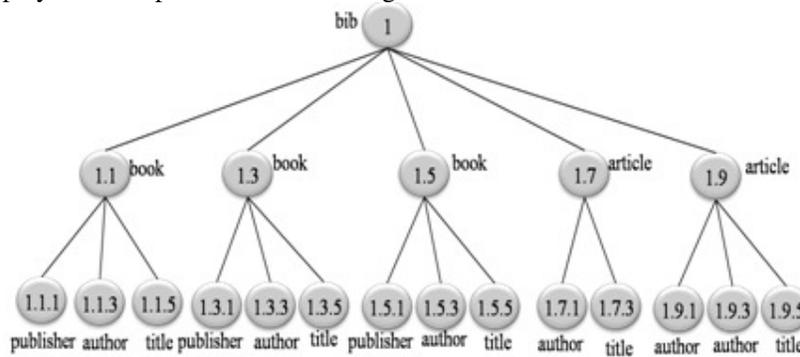


Figure 3. Simple XML document with ORDPATH labels

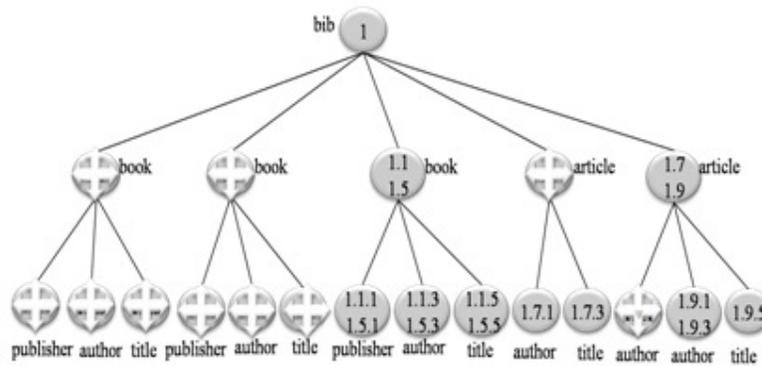


Figure 4. The compacted structure using CXDLS

4. BYTE REPRESENTATION OF THE LABELS

To achieve low storage consumption for XML documents, we have to reduce the size of node labels. Therefore, both ORDPATH and Cluster labeling schemes used Unicode-like compact representation that consists of a compressed binary representation and a prefix free encoding. It uses successive variable length L_i/O_i bitstrings and is generated to maintain document order and allow cheap and easy node comparisons. One L_i/O_i bitstring pair represents a component of a label. L_i bitstring specifies the number of bits of the succeeding O_i bitstring. The L_i bitstrings are represented using a prefix free encoding that can be constructed using a Huffman tree, an example for a prefix free encoding shown in figure 5(a). The binary encoding of a label is produced by locating each component value in the O_i value ranges and appending the corresponding L_i bitstring followed by the corresponding number of bits specifying the offset for the component value from the minimum O_i value within that range.

Example: Let us consider the bitstring pairs translation for the label (1.3.22). Note that the first component '1' is located in the O_i value range of [0, 7]. So that the corresponding L_0 bitstring is 01 and the length $L_0 = 3$, indicating a 3-bit O_0 bitstring. We therefore encode the component "1" with $L_0 = 01$ and $O_0 = 001$. Similar to that the binary encoding of the component "3" is the bitstring pair $L_1 = 01$, $O_1 = 011$. The component 22 is located in the O_i value range of [8,23] and its corresponding L_2 bitstring 100 and the length $L_2 = 4$. Thus the O_2 bitstring is 1111 that is the

offset of 15 from 8 specified in 4 bits. As final result the bitstring 01001010111001111 is the binary encoding of the cluster label (1.3.22).

Variations of prefix free encoding schemes can be created using the idea of Huffman trees, Figure 5 show different forms of prefix free encoding schemes.

Because the labels are binary encoded and stored in a byte array, in the case of use the codes in Figure 5(a) or the codes in Figure 5(b), the last byte may be incomplete. Therefore, it is padded on the right with zeros to end on an 8-bit boundary. This padding can lead to an increase in the storage requirements. For example, by using codes in Figure 5(a), the binary encoding of 1.9 is 010011000001 but its total length in bytes is 2 bytes and will be stored as the following bitstring 0100110000010000. Also by using codes in Figure 5(a), the label 1.9, for example, results in the bit sequence 0111100001, but it is padded by zeros to store it in 2 byte arrays as 0111100001000000 bitstring. In order to avoid padding with zeros, prefix free encoding scheme, shown in figure 5(c), was designed in a way that each division already observes byte boundaries.

Bitstring	Li	Oi value range
01	3	[0, 7]
100	14	[8, 23]
101	6	[24, 87]
1100	8	[88, 343]
1101	12	[344, 4439]
11100	16	[4440, 69975]
11101	32	[69976, 4.3×10 ⁹]
11110	48	[4.3×10 ⁹ , 2.8×10 ¹⁴]

(a)

Bitstring	Li	Oi value range
01	0	[1, 1]
10	1	[2, 3]
110	2	[4, 7]
1110	4	[8, 23]
11110	8	[24, 279]
111110	12	[280, 4375]
1111110	16	[4376, 69911]
11111110	20	[69912, 1118487]

(b)

Bitstring	Li	Oi value range
0	7	[1, 127]
10	14	[128, 16511]
110	21	[16512, 2113663]
1110	28	[2113664, 270549119]
1111	36	[270549120, ~ 237]

(c)

Figure 5. Variations of prefix free encoding schemes

5. THE IMPACTS OF PREFIX FREE ENCODING SCHEMES

In order to examine the impact of prefix free encoding schemes, mentioned in the previous section, on reducing the storage size needed to store the XML documents that are compacted using our approaches CXQU and CXDLS. We did experiment to measure and compare the storage requirements of our approaches and our cluster labeling scheme with other labeling schemes, such as OrdPath and Dewey [7,14]. In the experiment, each approach is suffixed with a number that refers to a prefix free encoding scheme, where number 1 refers to Figure 5(a) and so on respectively. We conducted our experiment using a variety of both synthetic and real datasets that covered a variety of sizes [5, 9, 10, 11, 12, 13, 16], application domains, and document characteristics. Table 1 displays the different structural properties of the used datasets.

Table 1. XML datasets used in the experiments

Datasets	File name	Topics	Size	No. of elements	Max depth
D1	Mondial	Geographical database	1,77MB	22423	6
D2	OT	Religion	3,32 MB	25317	6
D3	NT	Religion	0,99 MB	8577	6
D4	BOM	Religion	1,47 MB	7656	6
D5	XMark	XML benchmark	113 MB	1666315	12
D6	NCBI	Biological data	427,47 MB	2085385	5
D7	SwissPort	DB of protein sequences	112 MB	2977031	6
D8	Medline02n0378	Bibliography medicine science	120 MB	2790422	8
D9	medline02n0001	Bibliography medicine science	58,13 MB	1895193	8
D10	Part	TPC-H benchmark	6,02 MB	200001	4
D11	Lineitem	TPC-H benchmark	30,7 MB	1022976	4
D12	Customer	TPC-H benchmark	5,14 MB	135001	4
D13	Orders	TPC-H benchmark	5,12 MB	150001	4
D14	TOL	Organisms on Earth	5,36MB	80057	243

It is clearly visible from the results of the experiment in Figures 6, 7 and 8, that the use of third prefix free encoding scheme in our approaches made them more efficient in term of storage requirements for various XML data sets, when compared to other prefix free encoding schemes. These results confirm that the success rate of the use of our approaches (SCQX, CXDLS and cluster labeling scheme) is very high and they can dramatically reduce the storage requirements for almost all the datasets.

From result in Figure 8, it can be observed that the storage requirements, by using the approach CXDLS, are very small for the documents such as PART, Lineitem, Order and Customer because they have a regular structure and CXDLS focuses on compacting regular XML structures. At the same time the storage requirements are still relatively small for other documents that have either an irregular structure or less regular structure.

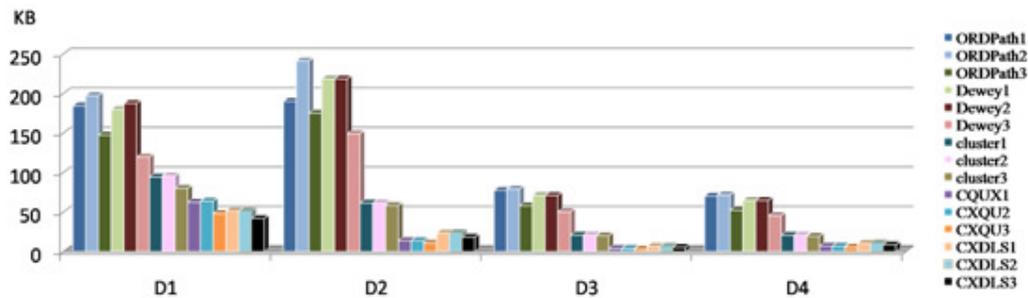


Figure 6. The storage requirements

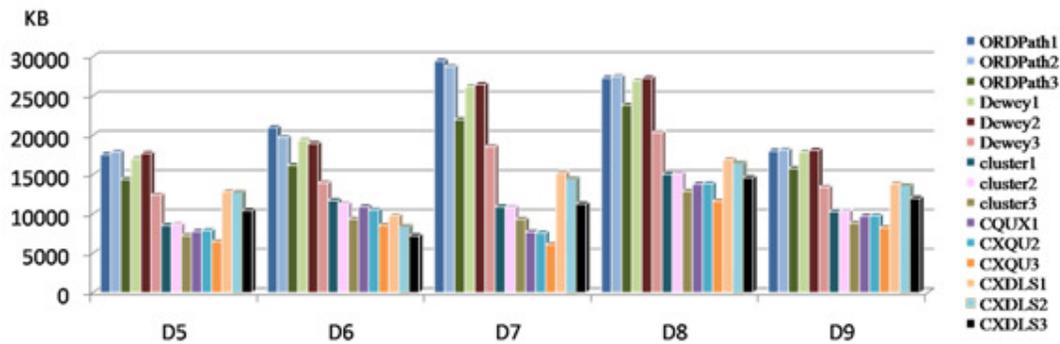


Figure 7. The storage requirements

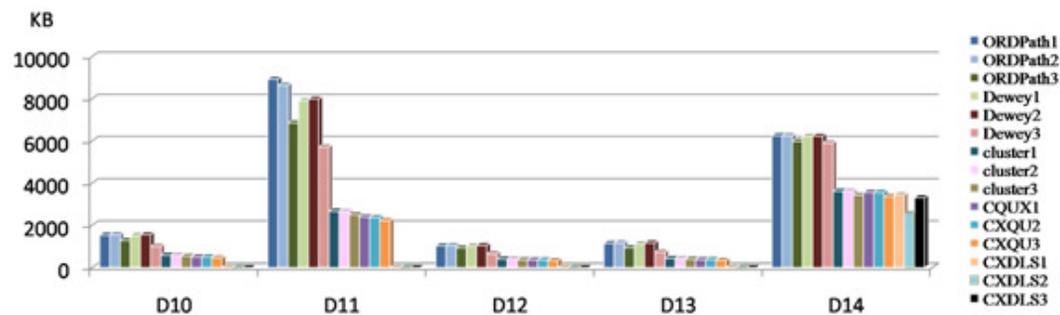


Figure 8. The storage requirements

6. CONCLUSIONS

This work investigated prefix free encoding technique created using the idea of Huffman trees. Our experimental results indicate that it is possible to provide significant benefits in terms of the storage requirements by using prefix free encoding, our compaction and labeling scheme techniques. An interesting future research direction is to explore more encoding formats and study how our compaction techniques could be extended to these formats. Since minimizing the storage costs can further improve query and update performance, one other possible future direction is to test the influence of prefix free encoding schemes on the query and update performance.

REFERENCES

- [1] R. Alkhatib and M. H. Scholl. Cxqu: A compact xml storage for efficient query and update processing. In P. Pichappan and A. Abraham, editors, ICDIM, pages 605–612. IEEE, 2008.
- [2] R. Alkhatib and M. H. Scholl. Compacting xml structures using a dynamic labeling scheme. In A. P. Sexton, editor, BNCOD, volume 5588 of Lecture Notes in Computer Science, pages 158–170. Springer, 2009.
- [3] M. Ali, M. A. Khan: Efficient parallel compression and decompression for large XML files. Int. Arab J. Inf. Technol. 13(4):403-408, 2016
- [4] H. AlZadjali, Siobhán North: XML Labels Compression using Prefix-encodings. WEBIST, pages 69-75, 2016
- [5] J. Bosak. Xml-tagged religion. Oct 1998. <http://xml.coverpages.org>.

- [6] S. Böttcher, R. Hartel, C. Krislin: CluX - Clustering XML Sub-trees. ICEIS, pages 142, 150, 2010
- [7] T. Härder, M. P. Haustein, C. Mathis, and M. W. 0002. Node labeling schemes for dynamic xml documents reconsidered. *Data Knowl. Eng.*, 60(1):126–149, 2007.
- [8] M. Lohrey, S. Maneth, R. Mennicke: XML tree structure compression using RePair. *Inf. Syst.* 38(8): 1150-1167, 2013
- [9] D. R. MADDISON, K.-S. SCHULZ, and W. P. MADDISON. The tree of life web project. *ZOOTAXA*, pages 19–40, 20 Nov. 2007.
- [10] W. May. Information extraction and integration with FLORID: The MONDIAL case study. Technical Report 131, Universität Freiburg, Institut für Informatik, 1999. Available from <http://dbis.informatik.uni-goettingen.de/Mondial>.
- [11] G. Miklau. Xml repository. <http://www.cs.washington.edu/research/xmldatasets>.
- [12] NCBI. National center for biotechnology information(ncbi) xml data format. <http://www.ncbi.nlm.nih.gov/index.html>.
- [13] NLM. National library of medicine (nlm) xml data format. <http://xml.coverpages.org>
- [14] P. E. O’Neil, E. J. O’Neil, S. Pal, I. Cseri, G. Schaller, and N. Westbury. Ordpaths: Insert-friendly xml node labels. In G. Weikum, A. C. König, and S. Deßloch, editors, *SIGMOD Conference*, pages 903–908. ACM, 2004.
- [15] Tatarinov, S. Viglas, K. S. Beyer, J. Shanmugasundaram, E. J. Shekita, and C. Zhang. Storing and querying ordered xml using a relational database system. In M. J. Franklin, B. Moon, and A. Ailamaki, editors, *SIGMOD Conference*, pages 204–215. ACM, 2002.
- [16] T. web project. the tol tree structure. 1998. <http://tolweb.org/tree>

USE OF ADAPTIVE COLOURED PETRI NETWORK IN SUPPORT OF DECISION-MAKING

Haroldo Issao Guibu¹ and João José Neto²

¹Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Brazil

²Escola Politécnica da Universidade de São Paulo, Brazil

ABSTRACT

This work presents the use of Adaptive Coloured Petri Net (ACPN) in support of decision making. ACPN is an extension of the Coloured Petri Net (CPN) that allows you to change the network topology. Usually, experts in a particular field can establish a set of rules for the proper functioning of a business or even a manufacturing process. On the other hand, it is possible that the same specialist has difficulty in incorporating this set of rules into a CPN that describes and follows the operation of the enterprise and, at the same time, adheres to the rules of good performance. To incorporate the rules of the expert into a CPN, the set of rules from the IF - THEN format to the extended adaptive decision table format is transformed into a set of rules that are dynamically incorporated to APN. The contribution of this paper is the use of ACPN to establish a method that allows the use of proven procedures in one area of knowledge (decision tables) in another area of knowledge (Petri nets and Workflows), making possible the adaptation of techniques and paving the way for new kind of analysis.

KEYWORDS

Adaptive Petri Nets, Coloured Petri Nets, Adaptive Decision Tables

1. INTRODUCTION

Coloured Petri Nets are an improvement of the original Petri Nets introduced by Carl Petri in the 1960s. Because of their ability to describe complex problems, their use has spread both in the engineering area and in the administrative area. Adaptive Coloured Petri Nets introduces an adaptive layer composed of several functions capable of changing the network topology, including or excluding places, transitions and arcs. In the area of decision support systems, the tools most used by specialists are the decision tables, which gave rise to several methods to help managers in their choices.

Among the methods developed for decision support there are the so-called multicriteria methods, which involve the adoption of multiple, hierarchically chained decision tables. In the process of improving the decision tables, new features are observed, which, although more complex, give the specialists the ability to describe their work model in a more realistic way. This paper describes the operation mode of the decision tables and the way of transcribing the rules of the tables for extended functions of Petri nets. By embedding a decision table in a Petri net, the simulation and analysis tools available in the Petri net development environments can be used, which leads to an increase in confidence in the decision criteria adopted.

Dhinaharan Nagamalai et al. (Eds) : CoSIT, SIGL, AIAPP, CYBI, CRIS, SEC, DMA - 2017

pp. 17– 27, 2017. © CS & IT-CSCP 2017

DOI : 10.5121/csit.2017.70403

2. DECISION TABLES

The Decision Table is an auxiliary tool in describing procedures for solving complex problems [9]. A Conventional Decision Table, presented in Table 1, can be considered as a problem composed of conditions, actions and rules where conditions are variables that must be evaluated for decision making, actions are the set of operations to be performed depending on the conditions at this moment, and the rules are the set of situations that are verified in response to the conditions .

. Table 1. Conventional Decision Tables.

	Rules column
Conditions rows	Condition values
Actions rows	Actions to be taken

A rule is constituted by the association of conditions and actions in a given column. The set of rule columns should cover all possibilities that may occur depending on the observed conditions and the actions to be taken. Depending on the current conditions of a problem, we look for which table rules satisfy these conditions:

- If no rule satisfies the conditions imposed, no action is taken;
- If only one rule applies, then the actions corresponding to the rule are executed;
- If more than one rule satisfies the conditions, then the actions corresponding to the rules are applied in parallel.
- Once the rules are applied, the table can be used again.
- The rules of a decision table are pre-defined and new rules can only be added or deleted by reviewing the table.

2.1. Adaptive Decision Tables

In 2001 Neto introduces the Adaptive Decision Table (ADT) [7] from a rule-driven adaptive device. In addition to rule lookup, an ADT allows you to include or exclude a rule from the rule set during device operation. As an example of its potential, Neto simulates an adaptive automaton to recognize sentences from context-dependent languages. In the ADT a conventional decision table is the underlying device to which a set of lines will be added to define the adaptive functions.

Adaptive functions constitute the adaptive layer of the adaptive device governed by rules. Modifying the rule set implies increasing the number of columns in the case of rule insertion, or decreasing the number of columns in the case of rule deletion. In both cases the amount of lines remains fixed. The Adaptive Decision Table (ADT) is capable to change its set of rules as a response to an external stimulus through the action of adaptive functions [7]. However, the implementation of more complex actions is not a simple task due to the limitation of the three elementary operations supported by ADT [9].

When a typical adaptive device is in operation and does not find applicable rules, it stops executing, indicating that this situation was not foreseen. For continuous operation devices, which do not have accepting or rejecting terminal states, stopping their execution would recognize an unforeseen situation and constitutes an error.

2.2. Extended Adaptive Decision Tables

To overcome this problem faced by continuous operation devices, Tchemra [9] created a variant of ADT and called it Extended Adaptive Decision Table (EADT), shown in Table 2

. Table 2. Extended Adaptive Decision Table.

		Adaptive Actions	Rules
Conventional Decision Table Set of auxiliary functions Adaptive layer	Criteria	Set of elementary	Criteria values
	Alternative		Actions to be applied
	Auxiliary functions	Adaptive actions	Auxiliary Functions to be called
	Adaptive functions		Adaptive actions to be performed

In EADT, adaptability does not apply only during the application of a rule, but also in the absence of applicable rules. A modifier helper device is queried and the solution produced by the modifier device is incorporated into the table in the form of a new rule, that is, in the repetition of the conditions that called the modifier device, the new rule will be executed and the modifier device will not need to be called.

3. PETRI NETS

3.1. Ordinary Petri Nets

Petri nets (PN) were created by Carl Petri in the 1960s to model communication between automata, which at that time also encompassed Discrete Event Systems (DES). Formally, a Petri net is a quadruple $PN = [P, T, I, O]$ where

P is a finite set of places;

T is a finite set of transitions;

I: $(P \times T) \rightarrow N$ is the input application, where N is the set of natural numbers;

O: $(T \times P) \rightarrow N$ is the output application, where N is the set of natural numbers

A marked network is a double $MPN = [PN, M]$, where PN is a Petri net and M is a set with the same dimension of P such that M (p) contains the number of marks or tokens of place p.

At the initial moment, M represents the initial marking of the MPN and it varies over time as the transitions succeed. In addition to the matrix form indicated in the formal definition of the Petri nets, it is possible to interpret the Petri nets as a graph with two types of nodes interconnected by arcs that presents a dynamic behaviour, and also as a system of rules of the type "condition \rightarrow action" which represent a knowledge base.

Figure 1 shows a Petri net in the form of a graph, in which the circles are the "Places", the rectangles are the "Transitions". The "Places" and the "Transitions" constitute the nodes of the graph and they are interconnected through the oriented arcs

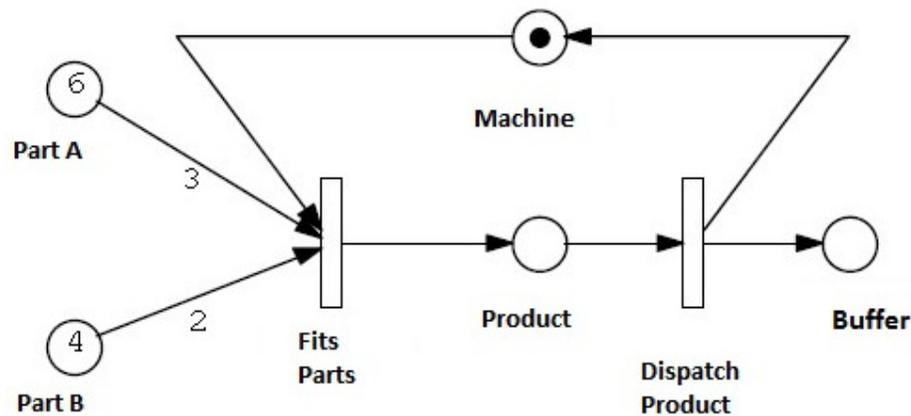


Figure 1. Example of a Petri Net

3.2. Reconfigurable Petri Nets

Several extensions of the Petri Nets were developed with the objective of simplifying the modelling of Discrete Events Systems, even for distributed environments. However, m Several extensions of the Petri Nets were developed with the objective of simplifying the modelling of Systems to Discrete Events, even for distributed environments. However, most extensions are not designed to model systems that change during their operation.

One group of Petri Nets extensions that attempts to tackle the problem of modelling systems that change during their operation is composed by the Self-Modifying Petri Nets [10], by the Reconfigurable Petri Nets via graph rewriting [6], by the Adaptive Petri Nets [1] and the Adaptive Fuzzy Petri Nets [5]. Each of these extensions has its own characteristics, but they share the fact that they can modify, during execution, the firing rules of the transitions or the topology of the network.

With the same name, the same acronym, but of different origins, we find in the literature Reconfigurable Petri Nets (RPN) introduced in [3] and in [6]. The work of Llorens and Oliver is an evolution of the work of Badouel and Oliver [1] and combines the techniques of graph grammars with the idea of Valk's Self-Modifying Petri Net, creating a system of rewriting of the network. In their work, Llorens and Oliver demonstrated the equivalence between the RPN and the PN in terms of properties and also that the RPN are equivalent to the Turing machines regarding the power of expression.

In Figure 2 we have schematized a Reconfigurable Petri Net according to Guan [3]. There are two interdependent layers, the control layer and the presentation layer. The places of the control layer are different in their nature from the places of the presentation layer.

Each place of the control layer has associated a set of functions that is capable to change the topology of the presentation layer, that is, they reconfigure the presentation layer. The tokens of the places are actually functions designed to modify the topology of the presentation layer.

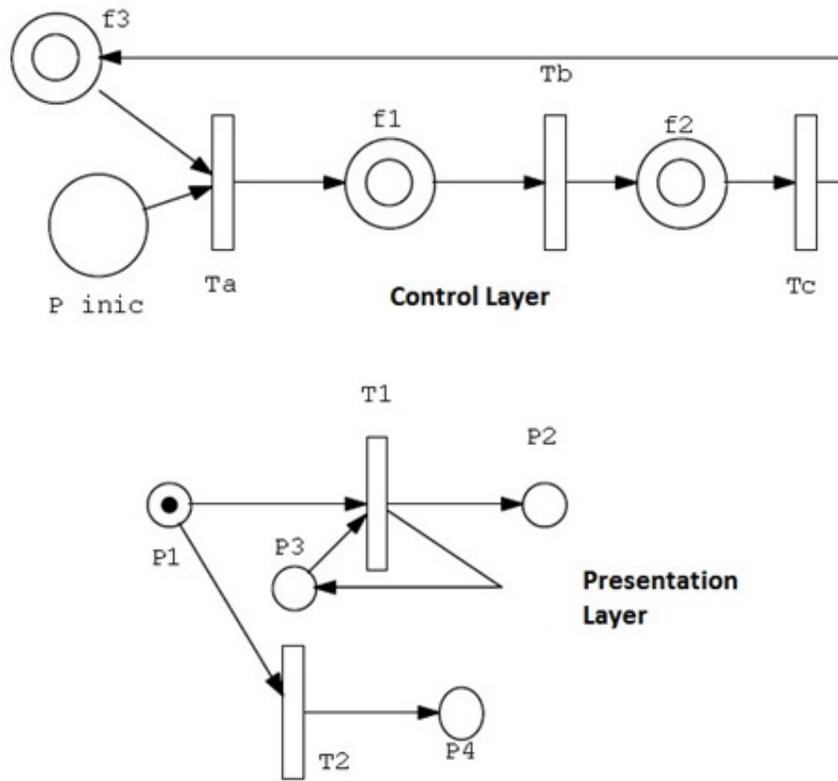


Figure 2. Reconfigurable Petri Net, version of Guan

3.3. Adaptive Petri Nets

An adaptive Petri net was defined by Camolesi [2] from the underlying device scheme plus adaptive layer as follows:

$APN = (C_0, AR_0, \Sigma, c_0, A, NA, BA, AA)$ in initial configuration c_0 .

Input stimuli move the APN to the next configuration if, and only if, a non-empty adaptive action is performed. In the k -th step we have:

$APN_k = (C_k, AR_k, \Sigma, c_k, A, NA, BA, AA)$, where

$APN = (PN_0, AM)$ is formed by an underlying initial device (PN_0) and an adaptive mechanism AM;

PN is the Petri Net device in step k . PN_0 is the initial underlying device and the set CR_0 represents the initial non-adaptive behaviour;

C_k is the set of all possible behaviours of PN in step k and $c_k \in C_k$ is its initial behavior in step k ;

ε ("empty string") denotes absence of valid element;

Σ is the set of all possible events that make up the input chain;

$A \subseteq C$ is the subset of PN acceptance configurations;

$F = C - A$ is the set of PN rejection configurations;

BA and AA are sets of adaptive actions, which include empty action;

$w = w_1 w_2 \dots w_n$ is the input string;

NA is a finite set of all symbols that can be generated as output by APN in response to the application of adaptive rules;

AR_k is the set of adaptive rules that define the adaptive behaviour of APN in step k and is given by a relation $AR_k \subseteq BA \times \Sigma \times C \times RP \times AA$.

AR_0 defines the initial behaviour of APN and the adaptive actions of insertion or elimination of places and transitions are transforming the set of rules.

The rules $reg \in AR_k$ are of the form

$\langle ba \rangle, (P, T, I, O), \langle aa \rangle$ and operate as follows:

A symbol $\sigma \in \Sigma$ causes reg to execute the action $ba \in BA$. If the action of $ba \in BA$ deletes reg from AR_k , the execution of reg is aborted, otherwise, the underlying rule of $reg = (P, T, I, O)$ applies. Finally, the adaptive action $aa \in AA$ is performed.

3.4. Adaptive Coloured Petri Nets

The Adaptive Coloured Petri Net uses the same scheme of wrapping the underlying device (CPN) with an adaptive layer (AL).

$ACPN = (CPN, AL)$ where

CPN is the conventional coloured Petri net,

LA = (AF, AR) is the adaptive layer.

In turn, the adaptive layer is composed by the set of adaptive functions (AF) and by the set of rules type IF - THEN (AR).

AF is the set of adaptive functions and is embedded in the Adaptive Coloured Petri net.

AR is the set of rules that must be inserted in the Adaptive Coloured Petri net through the execution of the adaptive functions.

The basic adaptive functions are inspection, insertion or incorporation and exclusion of a rule.

The ACPN uses the same strategy of RPN devised by Guan, but the control layer is a kind of interpreter of Decision Tables previously defined in order to produce the decision layer.

4. METHODOLOGY

The operation of the ACPN is based on an algorithm composed of four main phases:

Phase I: definition of the underlying decision table, with the inclusion of the criteria, alternatives and rules of the decision problem;

Phase II: generation of the decision matrix, whose values represent the relative weights and preferences of criteria and alternatives of the decision maker;

Phase III: transformation of the decision matrix in a XML format, in order to incorporate as a set of rules.

Phase IV: appending the XML file to the basic ACPN, resulting in the second layer, the decision layer.

The following example was adapted from [9] to illustrate the sequence of phases mentioned above, based in a decision procedure proposed in [8]. This is a decision problem in which the decision maker needs to certify and select suppliers of a particular product for his company. Two suppliers A and B are analyzed, according to the judgments of the decision-maker in accordance with selected for comparison:

C1 - Quality of services;

C2 - Flexibility for product delivery;

C3 - Distance from supplier and company location.

In phase I, criteria and alternatives to the problem are introduced in a conventional decision table, and the decision maker can create an initial set of rule, as showed in Table 3.

Table 3. Initial Decision Table.

		Rule1	Rule2	Rule3	Rule4
Criteria	C1 - Quality	Y	N	N	Y
	C2 - Flexibility	Y	Y	Y	Y
	C3 - Distance	Y	N	Y	N
Alternatives	A1 – Supplier A	X		X	X
	A2 – Supplier B		X		

In this example, the comparisons between the criteria are shown in Table 4, in which the decision-maker judges the criteria pairs, obtaining the matrix of reciprocal judgments[7].

Table 4. Judgement Matrix.

	C1	C2	C3
C1 - Quality	1	4	3
C2 - Flexibility	1/4	1	2
C3 - Distance	1/3	1/2	1

After checking the consistency of the values of judgment and normalization of values, the weights of each criterion are calculated, generating the vector of weights:

$$W = (0.62, 0.22, 0.16).$$

According to the judgment of the decision maker, the vector with the weight of each criterion represents the relative importance of each one. In this example, the resulting weights indicate that the criterion is more important in relative to others:

Quality: 0.62 - Flexibility: 0.22 -Distance: 0.16.

At this point, it is necessary to check the consistency of the criteria judgments. The matrix of comparison between the criteria of Table 4 is evaluated for the verification of the degree of consistency of the judgments, which is given by the consistency ratio (CR), as a function of the order of the matrix:

a) vector of weights $pe = (1.98, 0.69, 0.47)^T$

b) consistency vector $cs = (3.20, 3.08, 3.04)^T$

c) eigenvalue $\lambda_{max} = 3.11$

d) consistency index $CI = 0.05$

e) consistency ratio $CR = 0.09$

According to [8], the value of $CR = 0.09$ indicates that the judgments are consistent and acceptable since $CR < 10\%$, otherwise it would be necessary to review Table 4. The next operation is to obtain the performance matrix. For this, the alternatives are compared to the pairs, with each of the criteria. The comparisons made by the decision maker in the example are shown in Table 5.

Table 5. Pairwise comparison matrix.

	C1		C2		C3	
	A1	A2	A1	A2	A1	A2
A1 – Supplier A	1	8	1	6	1	4
A2 – Supplier B	1/8	1	1/6	1	1/4	1

Normalizing the matrices, we obtain the following values:

$$z_{1,1} = 0.89 \quad z_{1,2} = 0.86 \quad z_{1,3} = 0.80$$

$$z_{2,1} = 0.11 \quad z_{2,2} = 0.14 \quad z_{2,3} = 0.20$$

which are performance matrix cells.

Table 6. Performance matrix Z.

	C1	C2	C3
A1 – Supplier A	0.89	0.86	0.80
A2 – Supplier B	0.11	0.14	0.20

From the performance matrix Z we obtain vector ax containing the figures indicating the relative importance of the alternatives.

$\alpha x_1 = 0.85$ and $\alpha x_2 = 0.15$ indicating that in this example the alternative A1 is much better than the alternative A2 and supplier A must be chosen. Figure 3 shows the Petri Net version of the initial decision table.

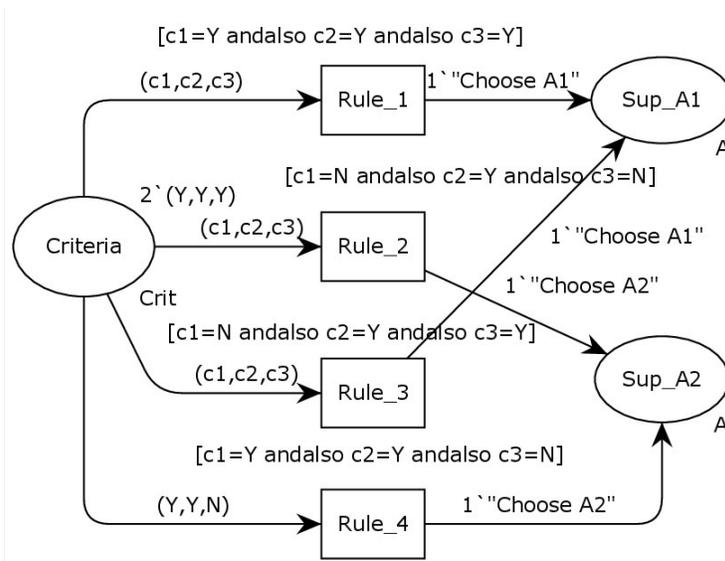


Figure 3. Petri Net equivalent of initial Decision Table

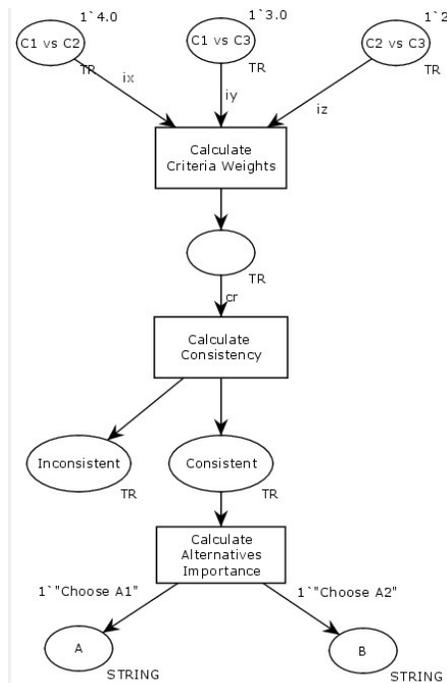


Figure 4. Petri Net Petri net of the decision-making process

Figure 4 summarizes the decision process in a graphical way.

5. CONCLUSIONS

In this paper we show how to incorporate decision tables in Petri nets. Well-established procedures in decision aid using decision tables can be used in a new tool that has additional analysis features to detect inconsistencies in procedures [4][13]. In daily activities, the business expert does not share his knowledge with the factory manager, losing synergy. Sharing two knowledges that are usually isolated provides a synergy. Good management practices are often disregarded in the day-to-day running of an enterprise because of the unawareness of the consequences of certain decisions, which are not always apparent.

In other areas where the use of Petri nets is widespread, gain is also possible, for example in flexible manufacturing systems [11][12]. Reconfigurable networks can be understood as a particular case of adaptive networks, where adaptation is achieved through reconfiguration of the network. The adaptive network is more general than the reconfigurable network because it can change its behavior while maintaining the same configuration by modifying the firing rules of the transitions.

In an adaptive network, the rules for operation in case of failures can be incorporated into the standard network, allowing greater agility in the operation without the need to stop the process until the experts in the failures take control. The recommendations of the experts would already be incorporated into the standard Petri net used in monitoring the operation. Reconfigurable systems during operation are a trend in the design of control systems and the ability to incorporate procedures from related areas is a feature that cannot be underestimated.

REFERENCES

- [1] Badouel E. ; Oliver, J. Reconfigurable Nets, a Class of High Level Petri Nets Supporting Dynamic Changes in Workflow Systems. (S.I.),1998.
- [2] Camolesi, A. R. Proposta de um Gerador de Ambientes para a Modelagem de Aplicações usando Tecnologia Adaptativa. Tese (Doutorado) —Escola Politécnica da Universidade de São Paulo, 2007.(in portuguese).
- [3] Guan S. U. ; Lim, S. S. Modeling adaptable multimedia and self-modifying protocol execution. Future Generation Computing Systems, vol.20, no 1, pp. 123-143, 2004.
- [4] Kheldoun, A., Barkaoui, K., Zhang, J., & Ioualalen, M. (2015, May). A High Level Net for Modeling and Analysis Reconfigurable Discrete Event Control Systems. In IFIP International Conference on Computer Science and its Applications_x000D_ (pp. 551-562). Springer International Publishing.
- [5] Little T. D. C.; Ghafoor, A. Synchronization and storage model for multimedia objects. IEEE J. Selected Areas Comm., pp. 413-427, Apr.1990., 1990.
- [6] Llorens, M. ; Oliver, J. Structural and dynamic changes in concurrent systems: Reconfigurable petri nets. IEEE Trans. Comput., vol. 53, no.9, pp. 1147-1158, 2004.
- [7] Neto, J. J. Adaptive rule-driven devices - general formulation and case study. Proc. 2001 Lecture Notes in Computer Science. Watson, B.W.and Wood, D. (Eds.): Implementation and Proc. 2001 Lecture Notes in Computer Science. Watson, B.W. and Wood, D. (Eds.): Implementation and Application of Automata 6th International Conf., Springer-Verlag,Vol.2494, pp. 234-250., 2001.
- [8] Saaty, T.L., "How to make a decision: the Analytic Hierarchy Process", Interfaces, Vol. 24, No. 6, pp19-43, 1994

- [9] Tchemra, A. H. Tabela de decisão adaptativa na tomada de decisões multicritério. Tese (Doutorado), Escola Politécnica da USP, São Paulo, 2009(in portuguese).
- [10] Valk, R. Self-modifying nets, a natural extension of petri nets. Lecture NotesLecture Notes in Com, vol. 62, pp. 464-476, 1978.
- [11] Zhang, J., Khalgui, M., Li, Z., Mosbahi, O., & Al-Ahmari, A. M. (2013). R-TNCES: A novel formalism for reconfigurable discrete event control systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43(4), 757-772.
- [12] Zhang, J., Khalgui, M., Li, Z., Frey, G., Mosbahi, O., & Salah, H. B. (2015). Reconfigurable coordination of distributed discrete event control systems. IEEE Transactions on Control Systems Technology, 23(1), 323-330.
- [13] Zhang, J. (2015). Modeling and verification of reconfigurable discrete event control systems (Doctoral dissertation, XIDIAN UNIVERSITY).

AUTHORS

Haroldo Issao Guibu

graduated in Electrical Engineering and MSc in Electrical Engineering in Polytechnic School of São Paulo (EPUSP) University, Brazil. Lecturer at Instituto Federal de Educação, Ciência e Tecnologia de São Paulo(IFSP) with main interest in Automation, PLC programming and Automation related adaptive technologies.



João José Neto

graduated in Electrical Engineering (1971), MSc in Electrical Engineering (1975) and doctor in Electrical Engineering (1980), and "livre docente" associate professor (1993) in the Polytechnic School of São Paulo (EPUSP) University, Brazil. Nowadays he is the head of LTA - Adaptive Technology Laboratory at the Department of Computer Engineering and Digital Systems at the EPUSP. His main experience is in the Computer Science area, with emphasis on the foundation of computer engineering and adaptivity. His main activities include adaptive devices, adaptive technology, adaptive automata and their applications to computer engineering and other areas, especially in adaptive decision making systems, natural language processing, compiler construction, robotics, computer education, intelligent system modeling, computer learning, pattern matching, inference and other applications founded on adaptivity and adaptive devices.



INTENTIONAL BLANK

A NOVEL ADAPTIVE - WAVELET BASED DETECTION ALGORITHM FOR CHIPLESS RFID SYSTEM

Meriam A. Bibile and Nemaï C. Karmakar

Department of Electrical and Computer Systems Engineering,
Monash University, Melbourne, Australia

ABSTRACT

In this paper, a novel wavelet-based detection algorithm is introduced for the detection of chipless RFID tags. The chipless RFID tag has a frequency signature which is identical to itself. Here a vector network analyser is used where the received backscatter signal is analysed in frequency domain. Thus the frequency signature is decoded by comparing the wavelet coefficients which identifies the bits accurately. Further, the detection algorithm has been applied for the tag detection under different dynamic environments to check the robustness of the detection algorithm. The new method doesn't rely on calibration tags and shows robust detection under different environments and movement.

KEYWORDS

Chipless RFID, wavelet, backscatter signal, frequency domain

1. INTRODUCTION

The chipless RFID reader extracts the backscattered signal and decodes the tag ID. This is an ongoing challenge, as the detection procedure for a chipless RFID tag has more complexities compared to a conventional RFID tag. The signal collides with other scatterers or tags which give a 'clutter' signal with interference. A number of detection techniques have been applied to achieve an accurate result of its tag ID.

The basic detection technique is based on comparing the received data with threshold values obtained by calibration. It is, therefore, a basic approach and it does not possess the flexibility and adaptability required in the detection process to address errors due to a dynamic environment [1]. Different types of detection algorithms and decoding techniques have been revealed in the past few years.

Moving average technique is a simple de-noising technique which removes noises by acting as a low pass filter. An 11 sample averaging moving average filtering has been successfully implemented on a low-cost mid-range microcontroller having low processing power capabilities, and a smoothed waveform is resulted after using this filtering technique. Hilbert transform

Dhinaharan Nagamalai et al. (Eds) : CoSIT, SIGL, AIAPP, CYBI, CRIS, SEC, DMA - 2017
pp. 29– 38, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70404

(HT) is a complex analytical signal processing technique [2]. This technique has been used to reconstruct the frequency signatures of the chipless tags. It has been experimentally proven that HT provides the extraction of the amplitude and phase functions of the frequency signature.

The signal space representation of chipless RFID tags uses an efficient mathematical model to decode information in a chipless RFID tag [3-4]. The frequency signatures are represented by a matrix which is composed of orthonormal column vectors and a singular value matrix. The constellation of signal points are plotted with a basis function. It can be seen that as the number of bits increase this method will face limitations. Matrix pencil method (MPM) and Short time matrix principle method (STMPM) are two more detection techniques that have been applied for chipless RFID systems [5-6]. These two techniques are applied in the time domain and are mentioned as accurate detection techniques in extracting the carrier to noise ratio (CNR) of the response. Detection is performed by extracting the poles and residues from the backscattered signal using the Matrix Pencil Algorithm. A Maximum Likelihood (ML) based tag detection technique and Trellis decoding technique has been developed where detection error rate is compared with the bit to bit detection [7]. It has been found that ML detection has the best performance. It also reports that the computational complexity is higher in ML detection technique than Trellis detection technique.

The main aim of this paper is to develop a detection algorithm which can be practically applied in firmware. As many of the above algorithms are highly complexed in implementing in the chipless RFID reader. Also, most of them have the limitation in the number of bits that can be detected. In this paper, we have developed a novel wavelet which suits the chipless RFID received signal to detect the frequency ID of the chipless RFID tag.

2. SYSTEM MODEL AND DESIGN

In this section the system model of the chipless RFID system used in this analysis is discussed. A description of the backscattered signal is also given and the backscattered signal of the used tag is analysed using simulation results obtained through CST microwave studio.

2.1. Experimental Model

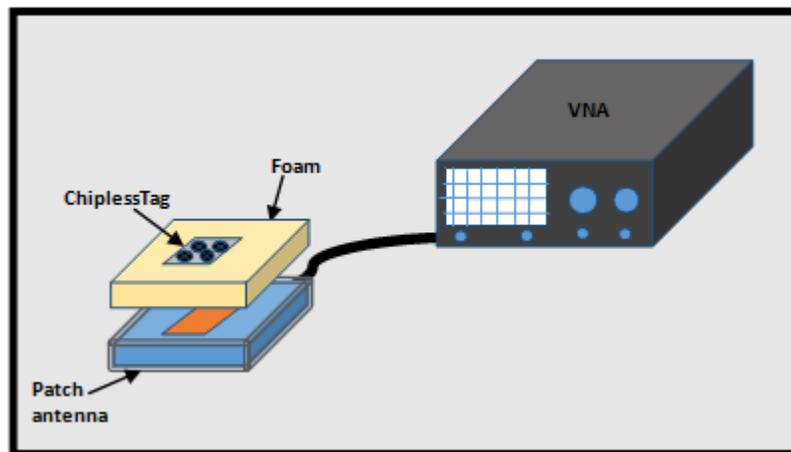


Figure 1 –Chipless RFID system

The experiment is performed using a 5-bit spiral resonator design. Each resonator has its own resonance frequency and has a diameter of 5mm. The tag with 5 bits has a dimension of 2cm × 2cm. The tag design is not presented in this paper to due to the confidentiality of the tag design. The tag RCS is shown in Fig 2. A patch antenna is used for the measurements, and the reading is taken by loading the antenna with a tag using vector network analyzer (VNA) as shown in Fig 1. The tag is loaded above the patch antenna and is placed on a piece of foam with 1cm thickness.

Backscattered signals from chipless RFID tags are very weak, and the detection process at the RFID reader is extremely susceptible to noise. This is because the information is represented as analog variations as opposed to a modulated digital stream of data as in conventional wireless communication of information where error correction schemes can be used to detect the presence of errors and discard erroneous data packets.

According to Fig 2, the resonant frequencies of the tag are located at 4.15, 4.70, 5.37, 6.10 and 6.95GHz respectively. The other local maxima detected in the signal are due to the amplitude spectrum of the background and antenna S11. These spurious peaks need to be carefully filtered to detect the correct frequency signature of the chipless RFID tag.

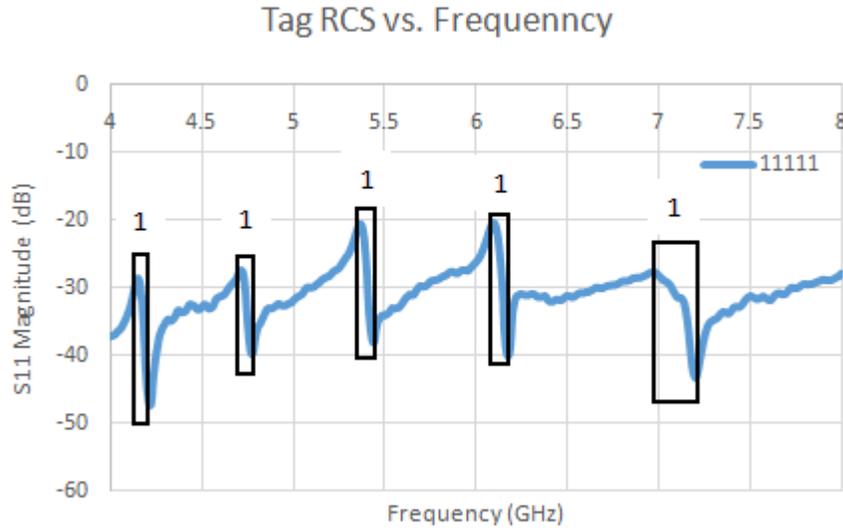


Figure 2 – S11 Magnitude (dB) vs. frequency of the 5-bit resonant tag

2.2. Wavelet Design

In this paper, a novel wavelet is adopted to detect the peaks of the backscattered signal. It is based on the Gaussian function which is given by (1).

$$f(x) = a \times e^{\frac{-(x-b)^2}{2c^2}} \quad (1)$$

where a is the height, b is the position and c is the width of the wavelet. The values a , b and c are adjusted to maximize the detection.

The novel wavelet design is shown in Fig 3. The width of the wavelet is adaptive according to the number of bits in the tag. In Fig 3, the width of the wavelet has been taken as 250MHz giving a

wider bandwidth to the wavelet. Further for the detection of the results this has been changed to 100MHz giving better resolution of the detected bits. This is an advantage of this algorithm as higher number of bits can be detected with better resolution.

The detected signal is compared with the wavelet, and a coefficient is defined to represent how closely correlated the wavelet is with each part of the signal. The larger the coefficient is in absolute value, the more the similarity appears. The wavelet is shifted until it covers the whole signal. The threshold coefficient is defined after performing the experiment for the tag with bit '11111' number of times under different environmental changes. By the variation in the detected signal under different circumstances a minimum and maximum coefficient is set giving rise to a confident band for detection.

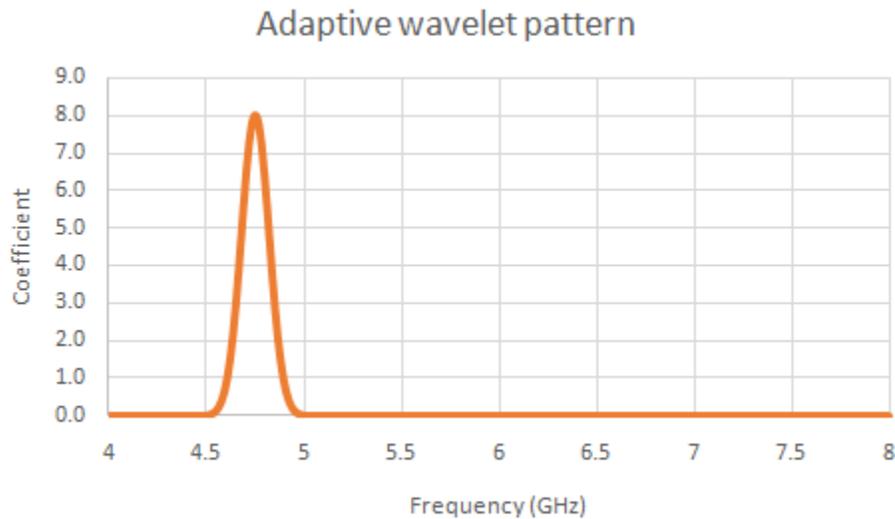


Figure 3 – New adopted wavelet pattern using Gaussian function

2.3. Flowchart for Detection Algorithm

The flowchart given in Fig 4, shows the steps of the developed algorithm and how it is applied to the received signal.

First the measured results using VNA are loaded into Matlab. The post processing is performed using Matlab programming software. The program can be directly downloaded to the microcontroller of the frequency domain reader developed by the research group which will be implemented in the next step of this research.

A baseline removal is performed on the first results to remove the nulls of the antenna S11. Then the wavelet is defined and is integrated with the resulting signal. The tag ID is decoded based on the condition band defined for the detection.

Experiments have been performed under different indoor dynamic environments such as placing clutter objects above the chipless RFID tag. Fig 8, shows the placing of hand 7cm above covering the tag adding back reflection to the received signal. Similarly, a copper plate was also placed above the tag at 7cm, and the received signal was observed using the VNA.

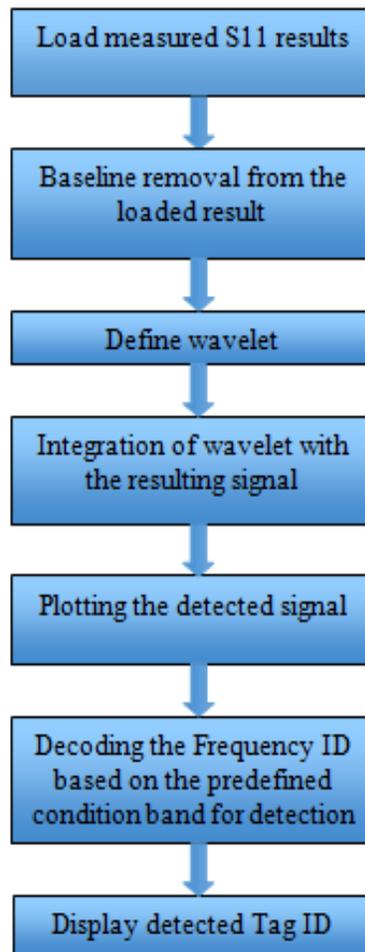


Figure 7 – Flowchart for applying the detection algorithm

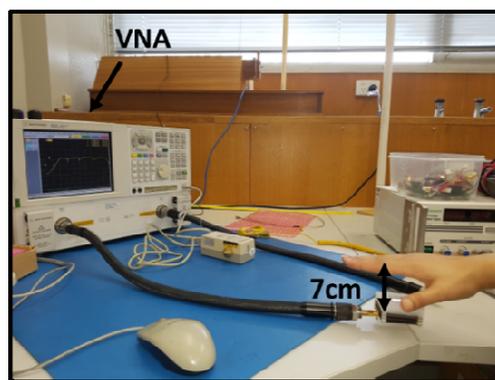


Figure 8 – placing hand on top of the tag at 7cm distance giving attenuation to the signal received at the VNA

3. RESULTS

The measured results of the tag detection are shown in Fig 9.

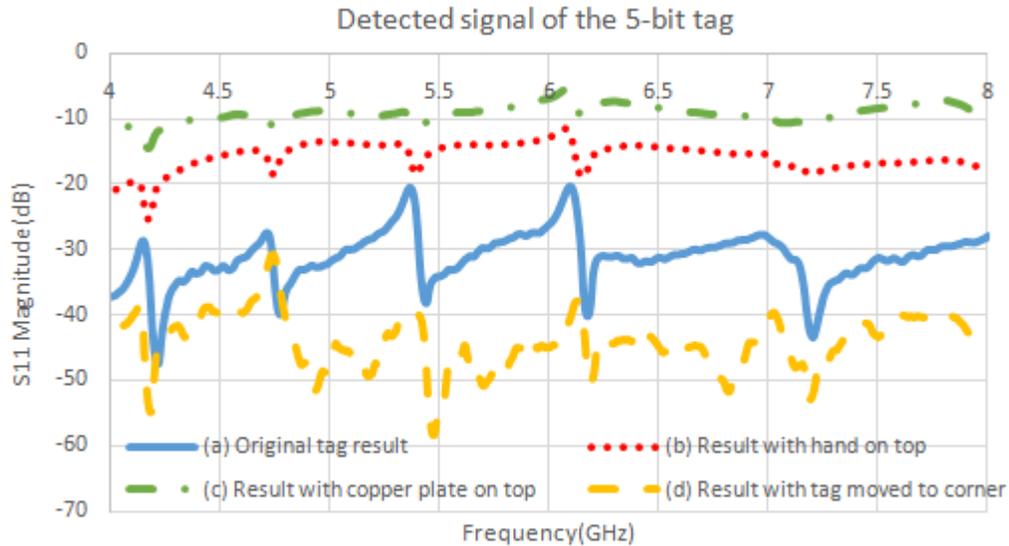


Figure 9 – Measured results (a) of the original tag at 1.5cm above the patch antenna, (b) with hand placed above the tag at a distance of 7cm, (c) with copper plate placed on top of the tag at a distance of 7cm, (d) by placing the tag at the corner of the patch antenna

It shows the received signals from the reader before applying the algorithm. In the original signal with the tag, we can identify the 5-bit resonances from the received signal itself. But when the hand, the copper plate is placed on top of the tag we can see that the received signal has increased in magnitude and has been shifted up. Also, peaks are not distinguishable at a glance. When the tag is placed in the corner of the antenna which implies the tag is moved out of the beam width of the antenna, we can again see a distortion in the signal as well as shifting the signal down in magnitude.

4. ANALYSIS

The detected signal after applying the detection algorithm is analysed in this section. The data is then interpreted to the end user. The results obtained after applying the developed wavelet based detection algorithm is given in Fig 10.

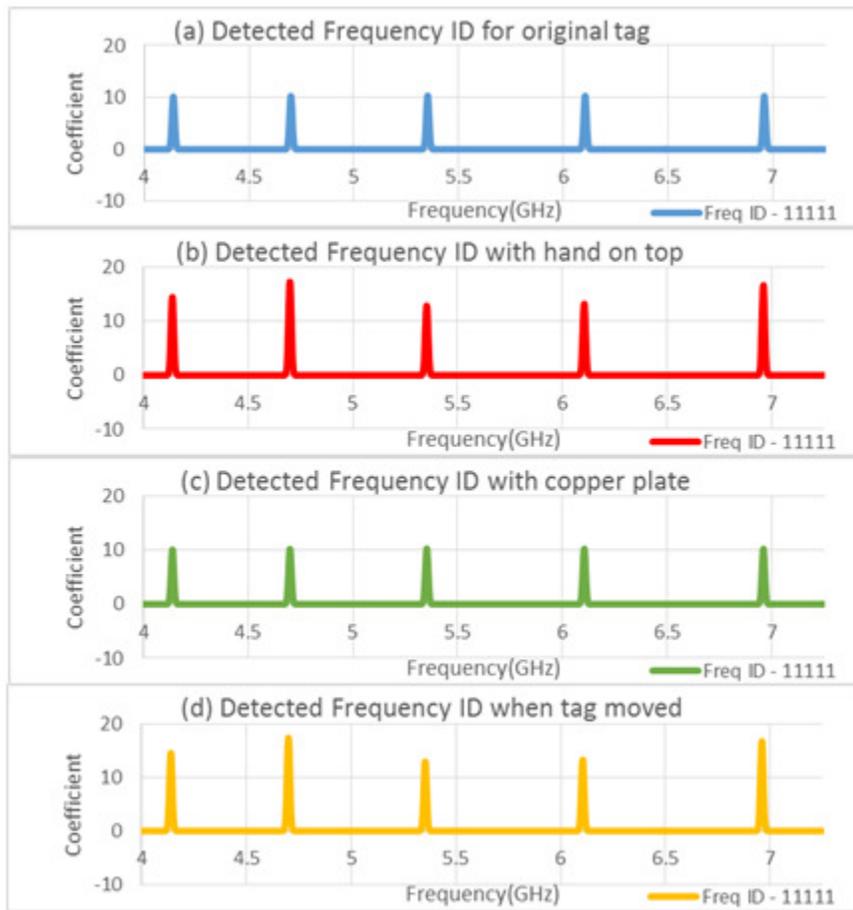


Figure 10 – Wavelet coefficients of the detected signals after applying the developed algorithm.
 (a) Original tag detection (b) Hand on top of tag (c) Copper plate on top of tag (d) Tag moved to the corner of antenna

The coefficients obtained are analysed in Table 1. The coefficients of the original tag are set to 10 which is taken as the threshold value of detection for the original tag with bits '11111'. A condition band is set for the maximum and minimum detectable coefficient bands. For the chipless RFID tag used in this experiment, the condition band is set between ± 8 , which gives +18 as the maximum value of detection as bit '1' and +2 as the minimum level of detection as bit '1'. If the coefficient is below or above these values, it will detect its ID as a bit '0'.

When the hand is placed, we can see that the coefficients are varying and the maximum coefficient reaches 17.35. When the copper plate is placed, it reaches a maximum value of 10.32 and the tag moved to the corner of the antenna gives a coefficient of 17.35. According to these results, the frequency ID of the tag is detectable as '11111' under all the applied conditions.

Table 1 – Wavelet coefficients for (a) Original tag detection – when the antenna is loaded with tag only (b) Hand on top of tag – when the hand is placed at 5cm distance above the tag (c) Copper plate on top of tag - when a copper plate of dimensions 10× 10× 0.2mm is placed at 5cm distance above the tag (d) Tag moved to the corner of antenna - when the tag is placed in the corner of the antenna

Resonance Freq. (GHz)	Original tag detection (Coefficient value)	Hand on top of tag	Copper plate on top of tag	Tag moved to corner of antenna
4.15	10.14	14.53	10.15	14.53
4.70	10.25	17.35	10.25	17.35
5.37	10.32	12.88	10.32	12.88
6.10	10.29	13.26	10.29	13.26
6.95	10.27	16.72	10.27	16.72

The frequency domain reader developed in MMARS lab uses the same patch antenna design for detection and is working in the frequency range of 4 – 7.5GHz. Therefore future direction of this research will be the implementation of the novel detection algorithm in firmware.

5. CONCLUSIONS

According to these results, it can be seen that a hand or conductive material (Cu) placed in the close vicinity of the tag (less than 10cm) will have an effect on the magnitude of the tag measurement. Still applying the novel detection algorithm the tag can be identified even if the conductive material is placed covering the tag. Also, when the tag is moved around the surface area covered by the antenna, the tag can be detected using this detection algorithm. Since the experiment was conducted by placing the clutter only on one side of the tag, more experiments should be performed by varying the position, size, material, etc. to conclude the probability of error due to the different material surrounding the tag.

An advantage of this detection algorithm is that it could detect up to a minimum of 1dB magnitude variation with added noise. The adaptability of the wavelet is another advantage of this algorithm as it can be applied for any type of tag in time domain or frequency domain and could cover the required frequency range. Also the detected bit resolution is high, which leaves room for the applications with higher number of bits in future.

When compared with other algorithms this has the most adaptability to any kind of chipless RFID system. Comparing the number of bits of a tag, the presented algorithm can be applied for higher number of bits whereas algorithms developed by Kalansuriya[3], Rezaiesarlak [6] and Divarathna[7] are implemented for low number of bits due to their mathematical complexity. As the width of the wavelet used in the detection is only 10MHz, with a guard band of another 10MHz a 20bit tag can be easily detected in the given frequency range of 4-8GHz.

The algorithm can be adapted into a wider bandwidth of frequencies and can be applied for any other chipless tag design in time or frequency domain.

Most of the developed detection algorithms in this research area are tested only for simulation results or for measured results using a vector network analyser. Presently the algorithm has been programmed into the microcontroller of the chipless RFID reader. Therefore this tag detection algorithm can be used further to investigate the robustness of the chipless RFID reader system under a dynamic environment. A further study and comparison of the detection error rate applying this algorithm will be performed based on results. It can be concluded that the simplicity of this algorithm allows it to be implemented in firmware and can be further fine-tuned to give robust detection of the tag in a real environment of chipless RFID tag detection.

ACKNOWLEDGEMENTS

This work is supported by the Australian Research Council Link Project Grant LP130101044: Discreet Reading of Printable Multi-bit Chipless RFID Tags on Polymer Banknotes.

REFERENCES

- [1] R. V. Koswatta, and N. C. Karmakar, "A novel reader architecture based on UWB chirp signal interrogation for multiresonator-based chipless RFID tag reading," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 9, pp. 2925-2933, 2012.
- [2] R. V. Koswatta, and N. C. Karmakar, "Moving Average Filtering Technique for Signal Processing in Digital Section of UWB Chipless RFID Reader," *Proceedings of Asia-Pacific Microwave Conference*, pp. 1304-1307, 2010.
- [3] P. Kalansuriya, N. C. Karmakar, and E. Viterbo, "Signal space representation of chipless RFID tag frequency signatures," *IEEE Globecom 2011 proceedings*, vol. 68, no. 7-8, pp. 437-445, 2013.
- [4] P. Kalansuriya, N. C. Karmakar, & Viterbo, E. (2013). On the detection of chipless RFID through signal space representation. *Annales des Telecommunications/Annals of Telecommunications*, 68(7-8), 437-445. doi: 10.1007/s12243-013-0377-4
- [5] R. Rezaiesarlak, & M. Manteghi, (2013). Short-time matrix pencil method for chipless RFID detection applications. *IEEE Transactions on Antennas and Propagation*, 61(5), 2801-2806. doi: 10.1109/TAP.2013.2238497
- [6] Rezaiesarlak, R., & Manteghi, M. (2014). A space-frequency technique for chipless RFID tag localization. *IEEE Transactions on Antennas and Propagation*, 62(11), 5790-5797. doi: 10.1109/TAP.2014.2350523
- [7] C. Divarathna, N.C. Karmakar, "A Maximum Likelihood Based Tag Detection Technique for MIMO Chipless RFID Systems," *IEEE International Microwave and RF Conference (IMaRC)*, 2014.
- [8] S. Preradovic, I. Balbin, N. C. Karmakar, and G. F. Swiegers, "Multiresonator-based chipless RFID system for low-cost item tracking," *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 5, pp. 1411-1419, May 2009.

- [9] A. Lazaro, A. Ramos, D. Girbau, and R. Villarino, "Chipless UWB RFID tag detection using continuous wavelet transform," *IEEE Antennas and Wireless Propag. Lett.*, vol. 10, pp. 520–523, May 2011.
- [10] A. Blischak and M. Manteghi, "Pole residue techniques for chipless RFID detection," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, Jun. 2009, pp. 1–4.

AUTHORS

Meriam Anushani Bibile (M'16) received the BSc. Degree in Engineering Physics from University of Colombo, Sri Lanka, in 2004 and MSc. Degree in Mobile, Personal and Satellite Communications from University of Westminster, London, United Kingdom, in 2006. She is currently working toward the Ph.D. degree in Electrical Engineering at Monash University, Melbourne, Australia. From 2008 – 2013 she was a Lecturer in Electronics and Telecommunications at the Institute of Technology, University of Moratuwa, Sri Lanka. Her areas of interests include chipless RFID, digital signal processing and digital electronics.



NemaiChandra Karmakar (S'91-M'91-SM'99) received the Ph.D. degree in information technology and electrical engineering from the University of Queensland, St. Lucia, Australia, in 1999. He has 20 years of teaching, design, and research experience in smart antennas, microwave active and passive circuits, and chipless RFIDs in both industry and academia in Australia, Canada, Singapore and Bangladesh. He has authored and co-authored more than 230 referred journal and conference papers, 24 referred book chapters and three edited and one co-authored books in the field of RFID. He has two patent applications for chipless RFIDs. Currently he is an Associate Professor with the Department of Electrical and Computer Systems Engineering, Monash University, Melbourne, Australia.



HANDWRITTEN CHARACTER RECOGNITION USING STRUCTURAL SHAPE DECOMPOSITION

Abdullah A. Al-Shaher¹ and Edwin R. Hancock²

¹Department of Computer and Information Systems, College of Business
Studies, Public Authority for Applied Education and Training, Kuwait

²Department of Computer Science, University of York, York, United Kingdom

ABSTRACT

This paper presents a statistical framework for recognising 2D shapes which are represented as an arrangement of curves or strokes. The approach is a hierarchical one which mixes geometric and symbolic information in a three-layer architecture. Each curve primitive is represented using a point-distribution model which describes how its shape varies over a set of training data. We assign stroke labels to the primitives and these indicate to which class they belong. Shapes are decomposed into an arrangement of primitives and the global shape representation has two components. The first of these is a second point distribution model that is used to represent the geometric arrangement of the curve centre-points. The second component is a string of stroke labels that represents the symbolic arrangement of strokes. Hence each shape can be represented by a set of centre-point deformation parameters and a dictionary of permissible stroke label configurations. The hierarchy is a two-level architecture in which the curve models reside at the nonterminal lower level of the tree. The top level represents the curve arrangements allowed by the dictionary of permissible stroke combinations. The aim in recognition is to minimise the cross entropy between the probability distributions for geometric alignment errors and curve label errors. We show how the stroke parameters, shape-alignment parameters and stroke labels may be recovered by applying the expectation maximization EM algorithm to the utility measure. We apply the resulting shape-recognition method to Arabic character recognition.

KEYWORDS

point distribution models, expectation maximization algorithm, discrete relaxation, hierarchical mixture of experts, Arabic scripts, handwritten characters

1. INTRODUCTION

The analysis and recognition of curved shapes has attracted considerable attention in the computer vision literature. Current work is nicely exemplified by point distribution models [1] and shape-contexts [2]. However, both of these methods are based on global shape-descriptors. This is potentially a limitation since a new model must be acquired for each class of shape and this is an inefficient process. An alternative and potentially more flexible route is to use a structural approach to the problem, in which shapes are decomposed into arrangements of

primitives. This idea was central to the work of Marr [3]. Shape learning may then be decomposed into a two-stage process. The first stage is to acquire a models of the variability is the distinct primitives. The second stage is to learn the arrangements of the primitives corresponding to different shape classes.

Although this structural approach has found widespread use in the character recognition community, it has not proved popular in the computer vision domain. The reason for this is that the segmentation of shapes into stable primitives has proved to be an elusive problem. Hence, there is a considerable literature on curve polygonalisation, segmentation and grouping. However, one of the reasons that the structural approach has proved fragile is that it has been approached using geometric rather than statistical methods. Hence, the models learned and the recognition results obtained are highly sensitive to segmentation error. In order to overcome these problems, in this paper we explore the use of probabilistic framework for recognition.

We focus on the problem of developing hierarchical shape models for handwritten Arabic characters. These characters are decomposed into concave or convex strokes. Our statistical learning architecture is reminiscent of the hierarchical mixture of experts algorithm. This is a variant of the expectation maximisation algorithm, which can deal with hierarchically structured models. The method was first introduced in 1994 by Jordan and Jacobs [4]. In its simplest form the method models data using a doubly nested mixture model. At the top layer the mixture is over a set of distinct object classes. This is sometimes referred to as the gating layer. At the lower level, the objects are represented as a mixture over object subclasses. These sub-classes feed into the gating later with predetermined weights. The parameters of the architecture reside in the sublayer, which is frequently represented using a Gaussian mixture model. The hierarchical mixture of experts algorithm provides a means of learning both the gating weights and the parameters of the Gaussian mixture model.

Here our structural decomposition of 2D shapes is a hierarchical one which mixes geometric and symbolic information. The hierarchy has a two-layer architecture. At the bottom layer we have strokes or curve primitives. These fall into different classes. For each class the curve primitive is represented using a point-distribution model [5] which describes how its shape varies over a set of training data. We assign stroke labels to the primitives to distinguish their class identity. At the top level of the hierarchy, shapes are represented as an arrangement of primitives. The representation of the arrangement of primitives has two components. The first of these is a second point distribution model that is used to represent how the arrangement of the primitive centre-points varies over the training data. The second component is a dictionary of configurations of stroke labels that represents the arrangements of strokes at a symbolic level. Recognition hence involves assigning stroke symbols to curves primitives, and recovering both stroke and shape deformation parameters. We present a probabilistic framework which can be for the purposes of shape-recognition in the hierarchy. We apply the resulting shape-recognition method to Arabic character recognition.

2. SHAPE REPRESENTATION

We are concerned with recognising a shape $W = \{\overset{p}{w}_1, \dots, \overset{p}{w}_p\}$ which consists of a set of p ordered but unlabelled landmark points with 2D co-ordinate vectors $\overset{p}{w}_1, \dots, \overset{p}{w}_p$. The shape is assumed to be segmented into a set of K non-overlapping strokes. Each stroke consists of a set

of consecutive landmark points. The set of points belonging to the stroke indexed k is S_k . For each stroke, we compute the mean position

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{i \in S_k} \mathbf{w}_i \quad (1)$$

The geometry of stroke arrangement is captured by the set of mean position vectors $C = \{\mathbf{c}_1, \dots, \mathbf{c}_k\}$.

Our hierarchical model of the characters uses both geometric and symbolic representations of the shapes. The models are constructed from training data. Each training pattern consists of a set of landmark points that are segmented into strokes. We commence by specifying the symbolic components of the representation. Each training pattern is assigned to shape class and each component stroke is assigned to stroke class. The set of shape-labels is Ω_c and the set of stroke labels is Ω_s . The symbolic structure of each shape is represented a permissible arrangement of stroke-labels. For shapes of class $\omega \in \Omega_c$ the permissible arrangement of strokes is denoted by

$$\Lambda_\omega = \langle \lambda_1^\omega, \lambda_2^\omega, \dots \rangle \quad (2)$$

We model the geometry of the strokes and stroke-centre arrangements using point distribution models. To capture the shape variations, we use training data. The data consists of a set of shapes which have been segmented into strokes. Let the t^{th} training pattern consist of the set of p landmark co-ordinate vectors $X^t = \{\mathbf{x}_1^t, \dots, \mathbf{x}_p^t\}$. Each training pattern is segmented into strokes. For the training pattern indexed t there are K_t strokes and the index set of the points belonging to the k^{th} stroke is S_k^t . To construct the point distribution model for the strokes and stroke-centre arrangements, we convert the point co-ordinates into long-vectors. For the training pattern indexed t , the long-vector of stroke centres is $X_t = ((\mathbf{c}_1^t)^T, (\mathbf{c}_2^t)^T, \dots, (\mathbf{c}_{L_t}^t)^T)^T$. Similarly for the stroke indexed k in the training pattern indexed t , the long-vector of co-ordinates is denoted by $z_{t,k}$. For examples shapes belonging to the class ω , to construct the stroke-centre point distribution model we need to first compute the mean long vector

$$Y_\omega = \frac{1}{|T_\omega|} \sum_{t \in T_\omega} X_t \quad (3)$$

where T_ω is the set of index patterns and the associated covariance matrix

$$\Sigma_\omega = \frac{1}{|T_\omega|} \sum_{t \in T_\omega} (X_t - Y_\omega)(X_t - Y_\omega)^T \quad (4)$$

The eigenmodes of the stroke-centre covariance matrix are used to construct the point-distribution model. First, the eigenvalues e of the stroke covariance matrix are found by solving the eigenvalue equation $|\Sigma_\omega - e^o I| = 0$ where I is the $2L \times 2L$ identity matrix. The eigen-vector ϕ_i corresponding to the eigenvalue e_i^o is found by solving the eigenvector equation $\Sigma \phi_i^o = e_i^o \phi_i^o$. According to Cootes and Taylor [6], the landmark points are allowed to undergo displacements relative to the mean-shape in directions defined by the eigenvectors of the covariance matrix Σ_ω . To compute the set of possible displacement directions, the M most significant eigenvectors are ordered according to the magnitudes of their corresponding eigenvalues to form the matrix of column-vectors $\Phi_\omega = (\phi_1^o | \phi_2^o | \dots | \phi_M^o)$, where $e_1^o, e_2^o, \dots, e_M^o$ is the order of the magnitudes of the eigenvectors. The landmark points are allowed to move in a direction which is a linear combination of the eigenvectors. The updated landmark positions are given by $\hat{X} = Y_\omega + \Phi_\omega \gamma_\omega$, where γ_ω is a vector of modal co-efficients. This vector represents the free-parameters of the global shape-model.

This procedure may be repeated to construct a point distribution model for each stroke class. The set of long vectors for strokes of class λ is $T_\lambda = \{Z_{t,k} | \lambda_k^t = \lambda\}$. The mean and covariance matrix for this set of long-vectors are denoted by Y_λ and Σ_λ and the associated modal matrix is Φ_λ . The point distribution model for the stroke landmark points is $\hat{Z} = Y_\lambda + \Phi_\lambda \gamma_\lambda$.

We have recently described how a mixture of point-distribution models may be fitted to samples of shapes. The method is based on the EM algorithm and can be used to learn point distribution models for both the stroke and shape classes in an unsupervised manner. We have used this method to learn the mean shapes and the modal matrices for the strokes. More details of the method are found in [7].

3. HIERARCHICAL ARCHITECTURE

With the stroke and shape point distribution models to hand, our recognition method proceeds in a hierarchical manner. To commence, we make maximum likelihood estimates of the best-fit parameters of each stroke-model to each set of stroke-points. The best-fit parameters γ_λ^k of the stroke-model with class-label λ to the set of points constituting the stroke indexed k is

$$\gamma_\lambda^k = \arg \max_{\gamma} p(z_k | \Phi_\lambda, \gamma) \quad (5)$$

We use the best-fit parameters to assign a label to each stroke. The label is that which has maximum a posteriori probability given the stroke parameters. The label assigned to the stroke indexed k is

$$l_k = \arg \max_{\lambda \in \Omega_s} P(l | z_k, \gamma_\lambda, \Phi_\lambda) \quad (6)$$

In practice, we assume that the fit error residuals follow a Gaussian distribution. As a result, the class label is that associated with the minimum squared error. This process is repeated for each stroke in turn. The class identity of the set of strokes is summarised the string of assigned stroke-labels

$$L = \langle l_1, l_2, \dots \rangle \quad (7)$$

Hence, the input layer is initialised using maximum likelihood stroke parameters and maximum a posteriori probability stroke labels.

The shape-layer takes this information as input. The goal of computation in this second layer is to refine the configuration of stroke labels using global constraints on the arrangement of strokes to form consistent shapes. The constraints come from both geometric and symbolic sources. The geometric constraints are provided by the fit of a stroke-centre point distribution model. The symbolic constraints are provide by a dictionary of permissible stroke-label strings for different shapes.

The parameters of the stroke-centre point distribution model are found using the EM algorithm [8]. Here we borrow ideas from the hierarchical mixture of experts algorithm, and pose the recovery of parameters as that of maximising a gated expected log-likelihood function for the distribution of stroke-centre alignment errors $p(X | \Phi_\omega, \Gamma_\omega)$. The likelihood function is gated by two sets of probabilities. The first of these are the a posteriori probabilities $P(\lambda_k^\omega | z_k, \gamma_{\lambda_k^\omega}, \Phi_{\lambda_k^\omega})$ of the individual strokes. The second are the conditional probabilities $P(L | \Lambda_\omega)$ of the assigned stroke-label string given the dictionary of permissible configurations for shapes of class ω . The expected log-likelihood function is given by

$$L = \sum_{\omega \in \Omega_c} P(L | \Lambda_\omega) \left\{ \prod_k P(\lambda_k^\omega | z_k, \gamma_{\lambda_k^\omega}, \Phi_{\lambda_k^\omega}) \right\} \ln p(X | \Phi_\omega, \Gamma_\omega) \quad (8)$$

The optimal set of stroke-centre alignment parameters satisfies the condition

$$\Gamma_\omega^* = \arg \max_{\Gamma} P(L | \Lambda_\omega) \left\{ \prod_k P(\lambda_k^\omega | z_k, \gamma_{\lambda_k^\omega}, \Phi_{\lambda_k^\omega}) \right\} \ln p(X | \Phi_\omega, \Gamma_\omega) \quad (9)$$

From the maximum likelihood alignment parameters we identify the shape-class of maximum *a posteriori* probability. The class is the one for which

$$\omega^* = \arg \max_{\omega \in \Omega_c} P(\omega | X, \Phi_\omega, \Gamma_\omega^*) \quad (10)$$

The class identity of the maximum *a posteriori* probability shape is passed back to the stroke-layer of the architecture. The stroke labels can then be refined in the light of the consistent assignments for the stroke-label configuration associated with the shape-class ω .

$$l_k = \arg \max_{\lambda \in \Omega_s} P(\lambda | z_k, \gamma_l^k, \Phi_\lambda) P(L(\lambda, k) | \Lambda_\omega) \quad (11)$$

Finally, the maximum likelihood parameters for the strokes are refined

$$\gamma_k = \arg \max_{\gamma} p(\xi_k | \Phi_{l_k}, \gamma, \Gamma_\omega^*) \quad (12)$$

These labels are passed to the shape-layer and the process is iterated to convergence.

4. MODELS

In this section we describe the probability distributions used to model the point-distribution alignment process and the symbol assignment process.

4.1 Point-Set Alignment

To develop a useful alignment algorithm we require a model for the measurement process. Here we assume that the observed position vectors, i.e. \mathbf{w}_i are derived from the model points through a Gaussian error process. According to our Gaussian model of the alignment errors,

$$p(\xi_k | \Phi_\lambda, \gamma_\lambda) = 12\pi\sigma \exp[-12\sigma^2 (\xi_k - Y_\omega - \Phi_\omega \gamma_\lambda)^T (\xi_k - Y_\omega - \Phi_\omega \gamma_\lambda)] \quad (13)$$

where σ^2 is the variance of the point-position errors which for simplicity are assumed to be isotropic. The maximum likelihood parameter vector is given by

$$\gamma_\lambda^k = \frac{1}{2} (\Phi_\omega^T \Phi_\omega)^{-1} (\Phi_\omega + \Phi_\omega^T) (\xi_k - Y_\omega) \quad (14)$$

A similar procedure may be applied to estimate the parameters of the stroke centre point distribution model.

4.2 Label Assignment

The distribution of label errors is modelled using the method developed by Hancock and Kittler [9]. To measure the degree of error we measure the Hamming distance between the assigned string of labels L and the dictionary item Λ . The Hamming distance is given by

$$H(L, \Lambda_\omega) = \sum_{i=1}^K \delta_{l_i, \lambda_i^\omega} \quad (15)$$

where δ is the DiracDelta function. With the Hamming distance to hand, the probability of the assigned string of labels L given the dictionary item Λ is

$$P(L | \Lambda_\omega) = K_p \exp[-k_p H(L, \Lambda_\omega)] \quad (16)$$

$$\text{where } K_p = (1 - p)^K \text{ and } k_p = \ln \frac{1 - p}{p} \tag{17}$$

are constants determined by the label-error probability p .

5. EXPERIMENT

We have evaluated our approach on sets of Arabic characters. Figure 1 shows some of the data used for the purpose of learning and recognition. In total we use, 18 distinct classes of Arabic characters and for each class there are 200 samples. The landmarks are placed uniformly along the length of the characters. The top row shows the example pattern used for training a representative set of shape-classes. The second row of the figure shows the configuration of mean strokes for each class. Finally, the third row shows the stroke centre-points.

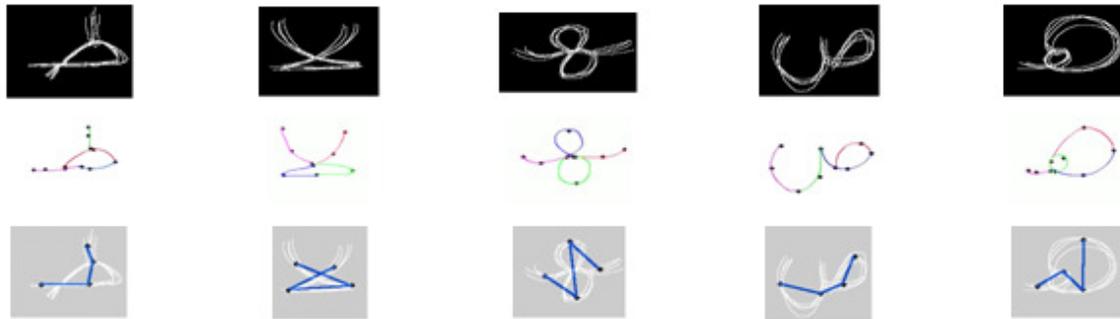


Figure 1: Row 1 shows sample training sets. Row 2 shows stroke mean shapes, Row 3 shows stroke arrangements

Table 1 shows the results for a series of recognition experiments. The top row of the table shows the character shape class. For each shape-class, we have used 200 test patterns to evaluate the recognition performance. These patterns are separate from the data used in training. The rows of the table compare the recognition results obtained using a global point-distribution model to represent the character shape and using the stroke decomposition model described in this paper. We list the number of correctly identified patterns. In the case of the global PDM, the average recognition accuracy is 93.2% over the different character classes. However, in the case of the stroke decomposition method the accuracy is 97%. Hence, the new method offers a performance gain of some 5%.

Table 1. Recognition Rate for shape-classes (Full Character, Stroke-based arrangements)

Shape												
Character	190	195	192	190	193	177	184	183	178	186	187	182
Stroke	197	198	197	199	198	189	196	192	194	190	193	188

Figure 2 examines the iterative qualities of the method. Here we plot the a posteriori class probability as a function of iteration number when recognition of characters of a specific class is attempted. The different curves are for different classes. The plot shows that the method converges in about six iterations and that the probabilities of the subdominant classes tend to zero.

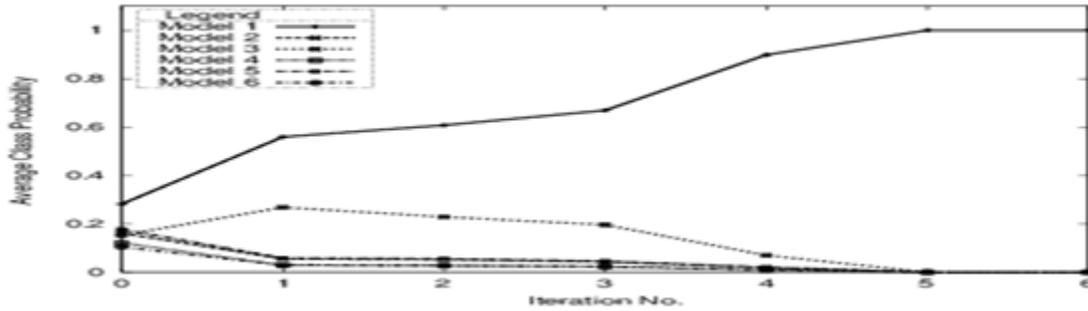


Figure 1. Alignment convergence rate as a function per iteration number

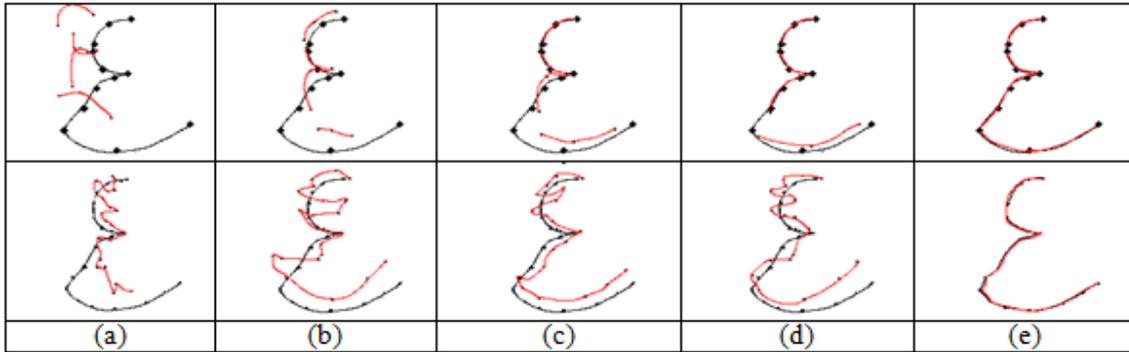


Figure 2. Alignment. First row shows hierarchical strokes:(a)iteration 1, (b)iteration 2, (c)iteration 3, (d)iteration 4, (e)iteration 5. Second row represents character models:(a)iteration 1, (b)iteration 2, (c)iteration 3, (d)iteration 4, (d)iteration 8

Figure 3 compares the fitting of the stroke model (top row) and a global PDM (bottom row) with iteration number. It is clear that the results obtained with the stroke model are better than those obtained with the global PDM, which develops erratic local deformations. Finally, we demonstrate in Figure 4 comparison of stroke decomposition and character with respect to recognition rate as a function of point position error. Stroke decomposition methods shows a better performance when points are moved randomly away of their original location.

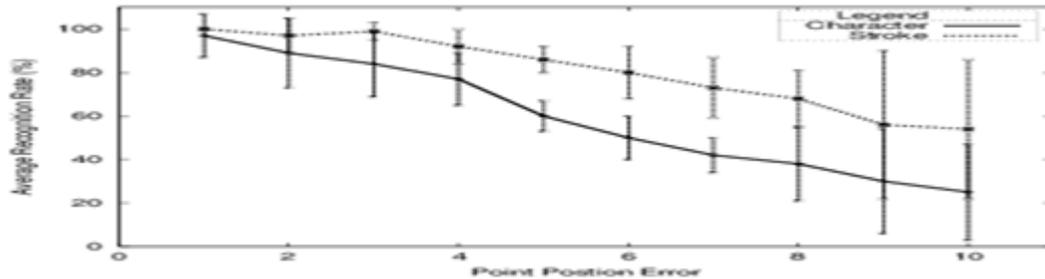


Figure 4. Recognition rate with respect to random point position

6. CONCLUSION

In This Paper, we have described a hierarchical probabilistic framework for shape recognition via stroke decomposition. The structural component of the method is represented using symbolic dictionaries, while geometric component is represented using Point Distribution Models. Experiments on Arabic character recognition reveal that the method offers advantages over a global PDM.

REFERENCES

- [1] T. Cootes, C. Taylor, D. Cooper and J. Graham, "Active Shape Models-Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-59, 1995.
- [2] S. Belongie, J. Malik and a. J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on PAMI*, vol. 24, no. 24, pp. 509-522, 2002.
- [3] D. C. Marr, *Vision: a computational investigation into the human representation and processing of visual information*, San Francisco: Freeman, 1982.
- [4] M. Jordan and R. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, pp. 181-214, 1994.
- [5] C. T. and T. C., "A mixture models for representing shape variation," *Image and Vision Computing*, vol. 17, pp. 403-409, 1999.
- [6] C. T. and T. C., "Combining point distribution models with shape models based on finite element analysis," *Image and Vision Computing*, vol. 13, no. 5, pp. 403-409, 1995.
- [7] A. A. Al-Shaher and E. R. Hancock, "Linear shape recognition with mixtures of point distribution models," in *SSPR*, Windsor, Canada, 2002.
- [8] A. Dempster, L. N. and R. D., "Maximum likelihood from incomplete data via the em algorithm," *Journal of Royal Statistical Soc. Ser.*, vol. 39, pp. 1-38, 1977.
- [9] E. R. Hancock and J. Kittler, "Discrete relaxation," *Pattern Recognition*, vol. 23, no. 7, pp. 711-733, 1990.

INTENTIONAL BLANK

DIVING PERFORMANCE ASSESSMENT BY MEANS OF VIDEO PROCESSING

Stefano Frassinelli, Alessandro Niccolai and Riccardo E. Zich

Dipartimento di Energia, Politecnico di Milano, Milan, Italy

ABSTRACT

The aim of this paper is to present a procedure for video analysis applied in an innovative way to diving performance assessment. Sport performance analysis is a trend that is growing exponentially for all level athletes. The technique here shown is based on two important requirements: flexibility and low cost. These two requirements lead to many problems in the video processing that have been faced and solved in this paper.

KEYWORDS

Diving performance, Video processing, Mosaicking, Colour filtering

1. INTRODUCTION

Many trends are involving in these years sports [1]: the number of people involved in fitness activities is increasing for every level and for every age [2], [3].

In this euphoria for sports, performance analysis is becoming every year much more important. Athletes of all levels are involved in analysis of their performances. This is giving to many sport a push toward higher and higher level of the competitions [4],[5].

Diving is one of the sports involved in this trend. The performances of diving are quite difficult to be objectively and engineeringly measured, but this is really important both for the athlete that can have an understanding of the efficiency of their training, both for competition organizers that can have a tool to make the judgment more objective.

Some wearable devices have been introduced in literature [6], but they can be only used during training because they are not allowed during competitions.

In other sports, like soccer [7], video analysis have been introduced with generic techniques and with ad-hoc procedures [8].

Some studies have been done also for diving [9], but the techniques proposed are quite expansive for many low-level athlete applications.

In this paper a low-cost procedure for performance assessment in diving by means of video processing will be explained with some proper example of the results of this procedure.

The paper is structured as follows: in Section 2 the performance metrics in diving will be briefly introduced. In Section 3 the needs and the requirements that have driven the development of this technique are shown. Section 4 shows the proposed procedure and eventually Section 5 contains the conclusions.

2. PERFORMANCE METRICS

Performance estimation can change a lot from one sport to another [10]. In some sports the important quantities are obvious, while in others they are hard to be identified clearly.

Diving is a sport in which the evaluation is based in a strong way on the subjective judgment of experts. In any case, it is possible to identify some measurable quantities that are related with the final performance of the dive.

Identifying the metrics is important because it gives a preferential path to the procedure to follow for the analysis of the dive.

The metrics here identified and studied are:

- Barycentre trajectory during the dive: this is an important metric [11] to understand possible mistakes made by the athlete during the jump, like lateral movements induced by a non-symmetric jump on the springboard [12], [13] or on the platform;
- Barycentre position during the entrance in the water: this is a metric that is important because it is one of the evaluation parameters that are commonly used in competitions;
- Maximum barycentre height [14]: this parameter is important because it gives an information of the time at disposal of the athlete for making all the figures that are required for the jump [15], [16].

These are not all the performance metrics that are used during competitions, but for sake of simplicity in this paper the analysis will be performed only of the barycentre positions. Other analysis can be done for examples on the forces transmitted to the platform or to the springboard. In the next sections the needs and the requirements that have driven the development of this technique will be explained.

3. NEED AND REQUIREMENTS

Needs and Requirements analysis is a fundamental step in technique designing [17]. This is useful to frame the range of applications and also to give a reasonable direction to the design phase. Moreover, it is important to finally assess the performance of the designed product, in this case the performance analysis technique.

The first, most important need is the flexibility of the technique: it should be used for almost any kind of video, for 10m platform diving and for 3m springboard diving.

These two possibilities are quite different because, while for 3m spring-board it is possible to imagine to have a fixed camera, for the 10m it is much more difficult: to have a good video it is necessary to place the camera far from the swimming pool and this leads to two problems: the first one is the space available. Again, for flexibility issues, the technique has to be suitable for the use in different places, so it must not require too much space. The second problem related is

the perspective problem: the diminution of the size of the diver can be a problem when it must be isolated from the background.

Another consequence of the flexibility requirement, is that the vide can be done also without a specific equipment, so it can have vibrations.

The second need, is the requirement of a technique that can be applied also for non-expert diver, so it must be cheap. It is not possible to imagine to have a high cost equipment.

Next section provides a general description of the overall procedure, then the main steps will be deeply analysed.

4. VIDEO PROCESSING PROCEDURE

Video processing is the technique applied here to performance assessment. While also other techniques can be applied [6], they are limited by regulatory issues in competitions and by comfort of the athlete.

Video analysis is an effective technique [9] because it does not require any additional item on the divers that can influence performance (the psychological approach is really important especially during high level competitions [12]) or that are forbidden[18].

It is important that the proposed technique can be applied both in training and during competitions: in this way, it is possible to compare the response of the athlete to the stress.

Another important advantage of video processing is that it can be applied to official video of competitions (Olympic Games or other International Competitions). These videos can be used as benchmark in the training.

Moreover, video processing results can be combined with kinematic simulations [19] to give to trainers and to athlete a complete tool in training.

In the next subsection, the overall process flow chart is explained; in the following ones, the steps are described providing some example of their application.

4.1. Process flow chart

The overall process, described in Figure 1, is composed by five steps and it is aimed to pass from a video, acquired during training or found on the Internet, to a performance score. In this paper the first four steps are described.

The first step is the image extraction by sampling correctly the video. To have a correct sampling, some parameters have to be chosen properly. Section 4.2 shows the concepts that are behind this choice.

The second and the third steps are the core of the procedure of image processing: firstly, a panorama is created by mosaicking, then the barycentre of the diver is found in each image by properly apply a colour filter. These two steps are described respectively in Section 4.3 and 4.4.

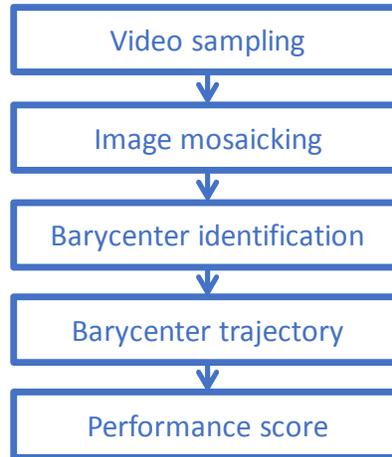


Figure 1. Process flow chart

Having found the barycentre of the diver it is possible to finally reconstruct his trajectory and finally to associate a score to the dive.

4.2. Video Sampling

The first step of the procedure is the extraction of the frames from the video. This procedure is easy because all the videos are defined as a sequence of frames. Only a free variable remains that is the sampling frequency.

This factor is defined by the compromise between calculation time and results precision.

Regarding the calculation time, it is possible to make some considerations: the diving time is defined by the height of the platform. In the case of a 10m platform, it is possible to firstly analyse a dive without any initial speed to have an idea of the total time:

$$\Delta t^2 = \frac{\Delta x}{2g} = 0.5 [s] \quad (1)$$

Where Δt is the diving time, Δx is the platform height and $g=9.81 \text{ m/s}^2$ is the gravity.

Considering that a dive can have almost three or four figures inside, each of them can last almost:

$$t_{fig} = \frac{\Delta t}{5} = 0.1 [s] \quad (2)$$

That correspond to an acquisition rate of:

$$f = \frac{1}{t_{fig}} = 10 [Hz] \quad (3)$$

To be sure of avoiding aliasing, it is necessary to acquire at least at 20 Hz. To have a safety margin an acquisition rate of 25Hz has been used.

This is the acquisition rate from the video (usually recorded with a higher acquisition rate) to the frame.

The safety margin in the acquisition rate is also useful to have a redundancy of information to be able to face to problems in video analysis.

Considering that the total video time is approximately between 2 and 5 second, the total number of frame that should be analysed is between 80 to 200 frames. This number is acceptable because it can be processed without problems with a commercial PC, so the requirement of non-specialist equipment is satisfied.

After having done a sampling from the video, it is possible to analyse firstly all the frames with the mosaicking procedure and then frame by frame by the barycentre identification.

4.3. Mosaicking

Having extracted from the video the frames, the second step of the procedure is mosaicking. Image mosaicking is a computational technique that exploit the presence of common features in a set of pictures [20] in images to a picture, called *panorama* [21].

Image mosaicking is a general [22] technique of image processing that has been applied to many fields, from medicine [23] to power plant inspection [24].

This technique is based on the identification of a set of features that are kept invariant during the most common transformations between pictures. The features of different pictures are matched and then the geometric transformation is identified. In this step to each picture a transfer function is associated [25].

At this point it is possible to wrap all the image in one image, called *panorama*. The *panorama* can be interpreted in this context as the common background to all the pictures.

In image mosaicking, it is possible to have different types of transformations [21]: testing many of them, it has been seen that the affine transformation is the best one for this application [26].

This step can solve two common problems related to the diving video: the first one is that the camera can have vibrations due to the low quality of the absence of an appropriate equipment. The second problem solved by mosaicking, is the fact that often, high quality video follows the diver, so the reference system changes (also a lot) from frame to frame.

1. Definition of the reference frame
2. Individuation of the main features of the reference frame
3. **For** all the frames **do**
4. Identification of the main feature of the frame
5. Feature matching between current frame and previous one
6. Definition of an appropriate geometric transformation
7. **End For**
8. Definition of the size of the panorama
9. Empty panorama creation
10. **For** all the frames **do**
11. Frame transformation in the global reference system
12. Transformed frame added to the panorama
13. **End For**

Algorithm 1. Mosaicking procedure

Both these two problems can be easily solved by the mosaicking technique: a panorama is created and to each frame a transfer function is associated.

The result of the application of image mosaicking on a diving non-professional video is shown in Figure 2. In this case the technique has been used to eliminate vibrations in the movie.



Figure 2. Example of image mosaicking for vibration elimination

In Figure 3 another example of application of the same procedure of image mosaicking has been applied on a video of a 10m platform dive of the Olympic Games of 2012. In this case the video has been recorded such that the athlete is almost always at the centre of the video.

After having defined a background common to all the frames, it is possible to go on with the procedure of barycentre identification

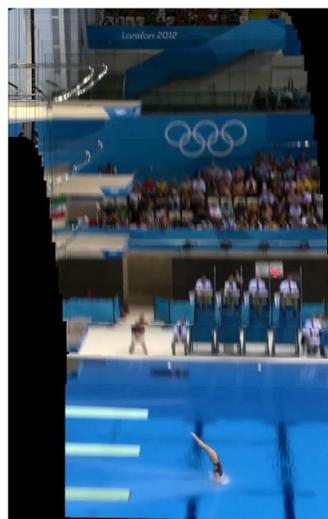


Figure 3. Example of image mosaicking for background reconstruction

4.4. Barycentre identification

Barycentre identification is a procedure that should be done frame by frame. In this paper, this procedure has been implemented using a proper colour filtering, as shown in Algorithm 2.

1. Filter parameters setting
2. Panorama filtering
3. **For** all the frames **do**
4. Definition of the frame in the global reference system
5. Frame colour filtering
6. Definition of the difference between the filtered panorama and the filtered frame
7. Object filtering
8. Barycentre calculation
9. **End For**

Algorithm 2. Barycentre identification procedure

It is possible to make some comments on the barycentre identification procedure:

- Due to the flexibility requirement, the colour filter [27] has to be not much selective because the light changes from frame to frame, so the colour of the diver can change. Obviously, a non-selective filter let many parts of the background in the filtered image. To reduce the number of false positive recognitions, also the background obtained from the mosaicking is filtered: in this way, the parts of the background that passes through the filter are known and they can be eliminated from the filtered frame;
- The colour filtering has been done by double threshold function applied to each channel in an appropriate colour space[28]. After several trials, the most suitable colour space is the HSV colour space [29].

Figure 4 shows two examples of the barycentre identification: the area output of the colour filter is the one with the white contour. The calculated barycentre is represented by the red square.

Analysing Figure 4 it is possible to see that, even if the filter is not perfect due to the presence of the swimsuit and due to the darker illumination of the arms of the diver, the position of the identified barycentre is approximately true.

Figure 5 shows the barycentre vertical position in time. It has been reconstructed doing the barycentre identification procedure for all the frames of the diving.

The results of the barycentre identification have to be processed because are affected by the noise introduced by little mistakes in the colour filtering procedure. To improve it, a moving average filter has been applied: in this way, the noise is reduced or completely eliminated.

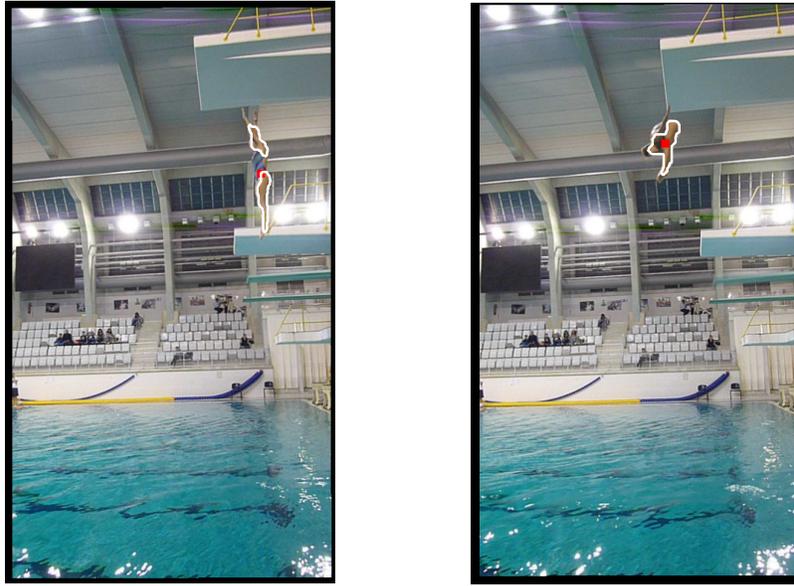


Figure 4. Example of barycentre identification

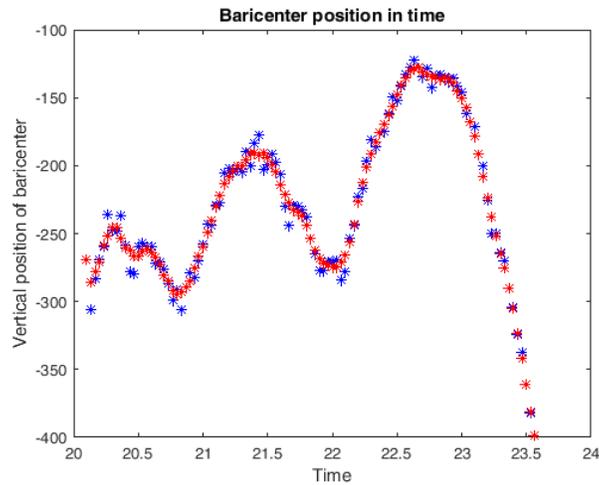


Figure 5. Barycentre vertical position in time. The blue dots are the positions directly taken from the analysis, while the red dots are the results of a filtering procedure

Figure 5 shows the comparison between the original position and the filtered ones: it is possible to notice that the filtered signal is able to reproduce correctly the large-scale movement of the athlete without introducing a delay. Moreover, the noise present in the original signal is correctly rejected.

5. CONCLUSIONS

In this paper a simple, economic and flexible procedure of video analysis of diving performance assessment has been presented. The technique is based on several steps that make the system as flexible as possible.

A possible improvement of the technique is the use of Deep Learning Neural Networks. Even if this last method is really powerful, it requires a huge number of in-put test cases: it is possible to apply the procedure described in this paper to prepare these inputs for the Network training.

REFERENCES

- [1] W. Thomson, "Worldwide Survey of Fitness Trends for 2015: What 's Driving the Market," *American College of Sports Medicine*, vol. 18, no. 6. pp. 8–17, 2014.
- [2] M. Stiefel, B. Knechtle, and R. Lepers, "Master triathletes have not reached limits in their Ironman triathlon performance," *Scand. J. Med. Sci. Sport.*, vol. 24, no. 1, pp. 89–97, 2014.
- [3] D. Sovak, M. R. Hawes, and K. Plant, "Morphological proportionality in elite age group North American divers," *J. Sports Sci.*, vol. 10, no. 5, pp. 451–465, 1992.
- [4] A. M. Shoak, B. Knechtle, P. Knechtle, A. C. Rüst, T. Rosemann, and R. Lepers, "Participation and performance trends in ultracycling," *Open Access J. Sport. Med.*, vol. 4, no. February, pp. 41–51, 2013.
- [5] R. Lepers, B. Knechtle, and P. J. Stapley, "Trends in Triathlon Performance: Effects of Sex and Age," *Sport. Med*, vol. 43, no. 9, pp. 851–863, 2013.
- [6] E. M. Kidman, "A Wearable Device with Inertial Motion Tracking and Vibro-tactile Feedback for Aesthetic Sport Athletes Diving Coach Monitor," 2016.
- [7] G. Hahm and K. Cho, "Event-based Sport Video Segmentation using Multimodal Analysis," 2016 Int. Conf. Inf. Commun. Technol. Converg., pp. 1119–1121, 2016.
- [8] A. Rehman and T. Saba, "Features extraction for soccer video semantic analysis: Current achievements and remaining issues," *Artif. Intell. Rev.*, vol. 41, no. 3, pp. 451–461, 2014.
- [9] L. Zhiwu, "Diving Sport Auxiliary Training Video Analysis Research," *Inf. Technol. J.*, vol. 13, no. 16, pp. 2514–2523, 2014.
- [10] S. Frassinelli, A. Niccolai, T. Marzi, M. Aghaei, M. Mussetta, and R. Zich, "Event-based measurement of power in sport activities by means of distributed wireless sensors," in *Proceedings of 1st International Conference on Event-Based Control, Communication and Signal Processing, EBCCSP 2015*, 2015.
- [11] P. J. Sinclair, C. A. Walker, and S. Cobley, "Variability and the control of rotation during springboard diving," in *32 International Conference of Biomechanics in Sports*, 2014, pp. 357–360.
- [12] S. Barris, D. Farrow, and K. Davids, "Increasing functional variability in the preparatory phase of the takeoff improves elite springboard diving performance," *Res. Q. Exerc. Sport*, vol. 85, no. 1, pp. 97–106, 2014.
- [13] P. W. Kong, "Hip extension during the come-out of multiple forward and inward pike somersaulting dives is controlled by eccentric contraction of the hip flexors," *J. Sports Sci.*, vol. 28, no. 5, pp. 537–543, 2010.
- [14] M. Sayyah and M. A. King, "Factors influencing variation in dive height in 1m springboard diving," 2016.
- [15] S. Barris, D. Farrow, and K. Davids, "Do the kinematics of a baulked take-off in springboard diving differ from those of a completed dive," *J. Sports Sci.*, vol. 31, no. 3, pp. 305–313, 2013.

- [16] M. R. Yeadon, "Twisting techniques used by competitive divers," *J. Sports Sci.*, vol. 11, no. 4, pp. 337–342, 1993.
- [17] M. Cantamessa, F. Montagna, and G. Cascini, "Design for innovation - A methodology to engineer the innovation diffusion into the development process," *Comput. Ind.*, vol. 75, pp. 46–57, 2016.
- [18] C. Walker, P. Sinclair, K. Graham, and S. Cobley, "The validation and application of Inertial Measurement Units to springboard diving," *Sport. Biomech.*, vol. 0, no. 0, pp. 1–16.
- [19] T. Heinen, M. Supej, and I. Čuk, "Performing a forward dive with 5.5 somersaults in platform diving: simulation of different technique variations," *Scand. J. Med. Sci. Sports*, pp. 1–9, 2016.
- [20] I. Aljarrah, A. Al-amareen, A. Idries, and O. Al-khaleel, "Image Mosaicing Using Binary Edge Detection," pp. 186–191, 2014.
- [21] D. Capel, "Image Mosaicing," in *Image Mosaicing and Super-resolution*, London: Springer London, 2004, pp. 47–79.
- [22] D. Ghosh and N. Kaabouch, "A survey on image mosaicing techniques," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 1–11, 2016.
- [23] H. Chen, "Mutual information based image registration for remote sensing data Mutual information based image registration for remote sensing data," 2003.
- [24] M. Aghaei, S. Leva, F. Grimaccia, and P. Milano, "PV Power Plant Inspection by Image Mosaicing Techniques For IR Real-Time Images," 2016 IEEE 43rd Photovolt. Spec. Conf., pp. 3100–3105, 2016.
- [25] A. Elibol, N. Gracias, and R. Garcia, "Fast topology estimation for image mosaicing using adaptive information thresholding," *Rob. Auton. Syst.*, vol. 61, no. 2, pp. 125–136, 2013.
- [26] R. M. S. Pir, "Image Mosaicing using MATLAB," vol. 3, no. 2, pp. 791–802, 2015.
- [27] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the HSV color space for image retrieval," *Int. Conf. Image Process.*, vol. 2, p. II-589-II-592, 2002.
- [28] J. Krauskopf, D. R. Williams, and D. W. Heeley, "Cardinal directions of color space," *Vision Res.*, vol. 22, no. 9, pp. 1123–1131, 1982.
- [29] K. Plataniotis and A. N. Venetsanopoulos, *Color image processing and applications*. Springer Science & Business Media, 2013.

IS AI IN JEOPARDY? THE NEED TO UNDER PROMISE AND OVER DELIVER – THE CASE FOR REALLY USEFUL MACHINE LEARNING

Martin Ciupa

CTO, calvIO Inc (part of the Calvary Robotics Group),
855 Publishers Pkwy, Webster, NY 14580, USA

ABSTRACT

There has been a dramatic increase in media interest in Artificial Intelligence (AI), in particular with regards to the promises and potential pitfalls of ongoing research, development and deployments. Recent news of success and failures are discussed. The existential opportunities and threats of extreme goals of AI (expressed in terms of Superintelligence/AGI and Socio-Economic impacts) are examined with regards to this media “frenzy”, and some comment and analysis provided. The application of the paper is in two parts, namely to first provide a review of this media coverage, and secondly to recommend project naming in AI with precise and realistic short term goals of achieving really useful machines, with specific smart components. An example of this is provided, namely the RUMLSM project, a novel AI/Machine Learning system proposed to resolve some of the known issues in bottom-up Deep Learning by Neural Networks, recognised by DARPA as the “Third Wave of AI.” An extensive, and up to date at the time of writing, Internet accessible reference set of supporting media articles is provided.

KEYWORDS

AI, Machine Learning, Robotics, Superintelligence, Third Wave of AI.

1. INTRODUCTION

In the past 18 months, though 2016 and into the first quarter of 2017 we are experiencing an amazing period of growth and media attention in Artificial Intelligence (AI) and Machine Learning (ML) [1]. Many high tech CEO’s have claimed association of their companies future regarding these technologies [2], massive investment is piling in [3]. While there have been many successes, there have been a few failures. The risk exists that “irrational exuberance” of the optimists may be a bubble that pessimists may be able to pop with claims that the AI vision is hyped [4]. The application of this research paper is to bring the current positive/negative media coverage to authors attention on this subject and to suggest a more precise, non-conflated, and modest use of terms in project terminology.

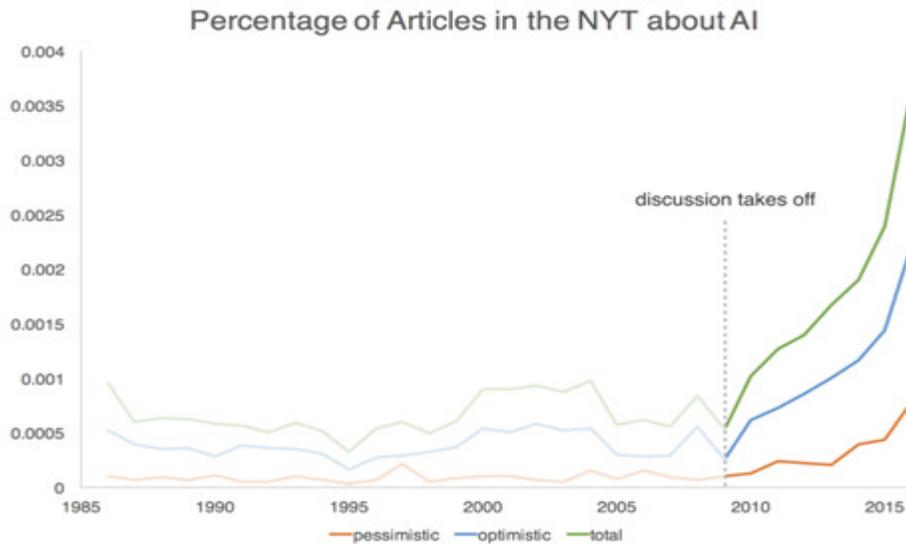


Figure 1. AI Media Coverage (Source [1])

For some commentators the very name “Artificial Intelligence” is controversial, implying something “unnatural.” The very word “Intelligence” defies definition in the minds of many [5]. For our purpose, we do not wish to address the philosophical artificiality question here. Rather, we will simply think of intelligence in working definition terms as: “that faculty of mind by which order is perceived in a situation hitherto considered to be disordered.” a definition (of Fatmi and Young) quoted in the Oxford Companion to the Mind [6]. Thus, if a machine, or an animal, has the facility to perceive the order change, then in a sense it has “intelligence”, and some aspect (perhaps limited, and with artifice) of “mind.” In this sense, it is a Cybernetic definition from a comparative systems perspective. We trust these points make the case that the AI term is itself controversial, it has its ideological proponents and detractors from the outset. In many respects, it would be useful to leave the term to the side and simply build really useful machines with specific descriptions of their application.

2. PROMISE – VISIONS OF AN EMERGING NEW POSITIVE PARADIGM

2.1. Recent success stories in AI and Machine Learning

There have been several notable success stories announced in the past 18 months, across a range of Machine Learning, Image Recognition, and Natural Language disciplines of AI. A high profile has been given to the ongoing development of Deep Learning by Reinforced/Convolutional Neural Network-based learning, in part underpinning IBM Watson/Jeopardy, AlphaGo and Poker game successes. These systems are beating the best human players who have invested their working lives into acquiring deep knowledge and intelligently applying it. See [7], [8] and [9]. The result of these successes has been a significant uptick in the media reporting of the AI potential and opportunities as mentioned earlier in [1], specifically more than four times as many New York Times articles discussed AI in 2016 than in 2009, as a percentage of the total number of articles published.

2.2. Investment into AI and Machine Learning

As a result of this positive news mentioned above, and the lead of many of the top tech company CEO's mentioned in [2], a significant financial round of investments has, and continues to be made, in this area. The market research company Forrester report that across all businesses, there will be a greater than 300% increase in investment in AI in 2017 compared with 2016 [3].

2.3. Timeline to AGI

The reporting and investment mentioned above have led to widespread speculation as to, "where is this all leading to?" And specifically, if the advent of "Superintelligent", Artificial Generic Intelligence (AGI) is on the foreseeable horizon. In 2015, the Machine Intelligence Research Institute compiled the MIRI AI predictions dataset, a collection of public predictions about human-level AGI timelines. Interesting features of the dataset include the result that the median dates at which people's predictions suggest AIG is less likely than not and more likely than not are 2033 and 2037 respectively. That is perhaps within 20 years. See Figure 2.

2.4. AI as the Singularity Hypothesis solution to Existential Risks

Several high profile scientists and philosophers have made their opinion on the issue mentioned above clear, optimistic and pessimistic, more about the pessimistic contributions will be mentioned in section 3 below. However, in the optimistic context of the promise of AI, and AGI in particular, much is suggested as opportunity to be sought by their agency. E.g., in their facility to help us to solve existential crises currently confronting humanity (such as Climate Change, Stabilisation of World Markets, Abundance of material needs, such as food, water and shelter, plus universal access to Healthcare and Education). The extreme optimistic positioning is that AGI, and the "Technology Singularity" will be humanity's last necessary invention, and once arrived a new age will ensure with material abundance for all, and solutions to humanity's problems [11].

2.5. Asilomar Conference 2017, World Economic Forum 2017

The Future of Life Institute, recognising the above opportunity, and in an attempt to offset negative pessimism about the downside of AGI, set out principles at the 2017 Asilomar conference, fundamentally a manifesto and guideline set of rules for the ethical development of AGI [12]. The explicit message is that action needs to be taken now to intercept any negativity and deliver positive result.

Similarly, the World Economic Forum, a high-profile international organisation whose stated mission is to "improve the state of the world", via Public-Private cooperation, continued on 17-20 January 2017 in its "exploration" on how developments in AI and Robotics could impact industry, governments and society in the future. Seeking to design innovative governance models to ensure that their benefits are maximised and the associated risks kept under control. The emphasis is that AI is coming, it is inevitable, and action has to be taken to ensure it arrives in good order [13].

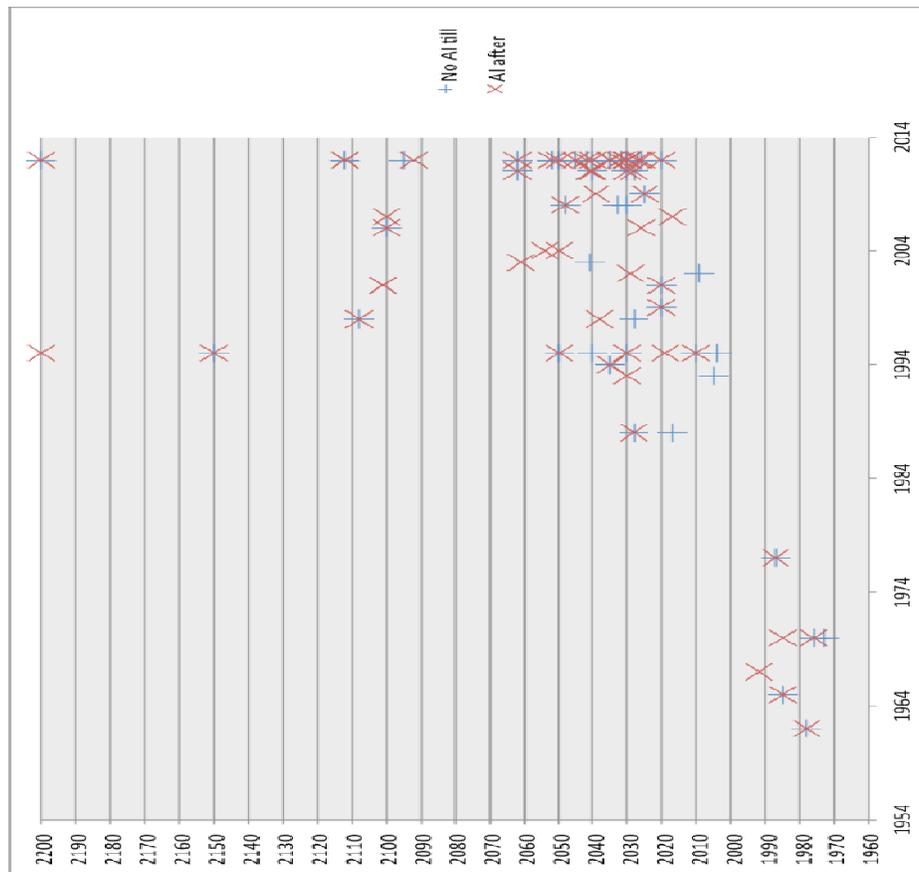


Figure 2: Predictions from the MIRI dataset surveys. Source [10]

2.6. RoboEthics

The academic study of Robot Ethics or “RoboEthics” is a well-established position in educational and research institutes. In part, this is a response to the understanding that robots, in particular drones, are entering the military as lethal weapons of war. And as such need regulation and legal codes of conduct [14]. But, is also concerned about the rights of robots themselves, as anticipated sapient and sentient “beings” with inalienable rights [15].

3. PITFALLS – RISK OF AI BEING OVERHYPED AND OR DANGEROUS

3.1. Recent failure stories in AI

But, not all voices are positive ones over AI. First of all, not all AI developments in the past few years have had positive results. Many no doubt go unreported, but two high profile ones in recent time are Microsoft’s Tay AI Bot (a Machine Learning online chatbot that learns from its interactions). Tay developed a “character” with strong racist, sexist opinions, and had to be removed [16] Furthermore, ongoing problems with AI messaging chatbots, at Google, resulted in a rollback in their adoption [17]. While these, perhaps overly ambitious projects did not meet the

expectations of the management, they nevertheless provide a proving ground for what can be done with today's technology and what needs further work. Such feedback is helpful in setting more realistic expectations [18].

There are other noted failures in AI projects, e.g., the fatal crash of a Tesla Automated Driving system in 2016, whereby autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky. In this case, it might be claimed that the testing/QA failed to identify this scenario and build sufficient responses to factor it safely.

3.2. Known problems with Deep Learning

Much, though not all of the recent success in AI has been due to Machine Learning advancement in so-called Deep Learning based Neural Networks. But, these bottom-up learning based systems (so-called because they develop learned experience from hierarchical analysis of basic data types and their deep correlations) have issues. They can learn well, often in tight domains beyond human learned experience and can outperform a human. But, a human can provide a top-down description, a discursive rationalisation to a, "why did you decide to make that move/act" question with a, "because, I think gaining control of the centre area a useful strategy at this point in the game" question. Whereas, a Deep Learning system cannot. It is not rule-based and cannot easily track its "reasoning". In a sense it is like an experienced financial markets trader who knows when to buy/sell, not because of the analytics, but because he has the intuitive "feel" built over years of trading. However, intuition is not necessarily a good guide when the market modalities change, as many traders have found out to their ultimate cost. As some critics have said, "it is only real intelligence if you can show it's working" [19].

In applications that require compliance with regulations and legal constraints, such systems may be considered risky. Furthermore, if such a system were committed to a highly Mission Critical Application, it would be difficult to question, supervise and control. Such intuitive systems, once relied upon, cannot easily be "rolled-back", human operators become reliant on the system to make all the hard decisions. In the event of failure, total system breakdown may ensue. There is a significant risk here. As AI is woven more deeply into the fabric of everyday life, the tension between human operators and AI has become increasingly salient. There is also a paradox: the same technologies that extend the intellectual powers of humans can displace them as well [20].

3.3. Mysterian Position – Beyond Computation

Some AI critics while recognising useful tools can be built do not believe AGI is likely in the short or medium term outlook of understanding of such things as "consciousness, free-will and self-aware visceral experience (qualia)." While this paper will not address these "Hard Problems of Consciousness" concerns; it is perhaps not necessary to incorporate these functional elements. If they cannot be defined in the case of natural human intelligence, we are not in a position to easily address them in our smart tools and AGI. Maybe AGI can pass Turing Tests, but not have these functions. The better question is perhaps, do we have them or are they illusory? See [21].

3.4. Issues of functional complexity and computational limits

Some AI critics, while not accepting necessarily the problems of intelligence being in principle beyond computational methods, believe it is so complex that our current computational

processing, is just not powerful enough, or the cost to provide it would be so astronomical as not to be worth the benefit. The above is a more of an objection in pragmatic terms. “Smart AI Tools” can be delivered in tight knowledge domains, but as for AGI, this is simply a non-starter for a long time to come. Recently Philosopher Daniel Dennett has made this argument, referring to AGI as “balderdash” [22]. Likewise, Philosopher Jerry Kaplan [23] has expressed concerns. However, some will point out that massive hierarchies of systems, even such as we have today, on parallel networked computational systems might cover the spectrum of human domain knowledge, with abstract reasoning and coordinated response. Such systems are termed Multi-Agent or Distributed AI systems [24]. Furthermore, such computer platforms, while not affordable today in their massive extension, might be within a few decades.

As for software, that also continues to improve, and the potential for geometric improvement by evolutionary learning algorithms, whereby AI works on AI and evolves rapidly is viable. The evolutionary cost of learning in simulated “toy universes” may not be as “expensive” or slow as in our experience of evolution in “natural reality.” Such simulations can be used for development, but also for QA/testing of extreme scenarios not easily tested physically [25].

3.5. Moore’s Law’s End?

While the hope is computer processing will continue to become dramatically cheaper, and more powerful, this is not a given. Moore’s Law that underpins that expectation is considered by some to be no longer applicable [26]. However, Intel counters that it is still alive [27]. The possibility of future improvement by innovation in new technology, such as Quantum Computers (and other innovative technologies) might step up and provide the substrates to keep Moore’s law alive for years to come [28].

3.6. “Prophets of Doom”, AGI as an Existential Risk

Even if the negative technical issues expressed above can be resolved, such that the promise and positive thinking expressed optimistically in section 2 above can be delivered, there are high profile voices being expressed that developing safe and ethical AGI within the next 20-30 years or so is very dangerous. These critics express concerns that AGI perhaps should not be developed. The danger is regarding the existential risk that they present to humanity, as well as to human culture regarding reducing human employment to such a point as to destabilise society.

Physicist Stephen Hawking, Microsoft founder Bill Gates and Tesla/SpaceX founder Elon Musk have expressed concerns about the possibility that AI could develop to the point that humans could not control it [29]. Stephen Hawking said in 2014 that "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." Hawking believes that in the coming decades, AI could offer "incalculable benefits and risks" such as "technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand." He makes the point that "The development of full artificial intelligence could spell the end of the human race" [30].

In January 2015, Nick Bostrom joined Stephen Hawking, Max Tegmark, Elon Musk, Lord Martin Rees, Jaan Tallinn, and numerous AI researchers, in signing the Future of Life Institute's open letter speaking to the potential risks and benefits associated with artificial intelligence. The

signatories "...believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today [31]. Their position is not that AGI should be halted, rather that urgent attention is needed to ensure that it is delivered ethically and safely. It should be noted that in the case of a sudden AGI "intelligence explosion", effective precautions will be extremely difficult. Not only would its creators have little ability to test their precautions on an intermediate intelligence, but the creators might not even have made any precautions at all, if the advent of the intelligence explosion catches them completely by surprise. Hence their concern to think through the AGI negative issues as early as possible so as to prepare principles to handle them.

3.7. AI Impact on Jobs – Luddite Rebellion

Furthermore, some consider AGI and the uptake of pervasive Robotics/Industrial Automation as harmful to our civilisation regarding reducing human employment to such a point as to destabilise society. Such instabilities in earlier Industrial Revolutions resulted in an anti-technology reaction, known the Luddite Rebellion [32]. Could this occur again? This concern needs to be addressed. Indeed it is the subject and concern of the World Economic Forum as mentioned earlier [13].

4. NEED TO UNDER PROMISE AND OVER DELIVER – AVOID THE HYPE

4.1. Addressing concerns

As Management Consultant Tom Peters says, "Quality is important, to be sure, so is absolute response time, and price, but at the top of most lists, by far, is keeping your word." With uncertainty rising, if you 'under promise, over deliver,' you will not only keep the customers satisfied; you'll keep the customers" [33]. While you can make the case this advice is not universally applicable, in cases where timelines are in doubt, and where there are critics in the wings who are willing to take pot shots, and your case depends, in part, on technology innovation not yet delivered, then it seems prudent to apply this principle.

With that in mind, I propose that we be more modest in our claims for AI and AGI. Yes, we may expect it to come, and indeed we may need it to help us solve knotty problems we face in an ever more chaotic and complex world full of existential problems, many arising out of our mismanagement. But, we should be wary of over-extending the use of the term, or conflating them. Thus, let us make a point of not talking up AGI for the next five years or so, to deliver on our projects in hand. That, if done well and to expectations will establish AI as a real vehicle for a paradigm change (and build solid foundations for AGI). As we have seen AI is an easily challenged term and easily hyped, we need to be more specific with our language, else "AI" risks becoming meaningless [34]. We should address specifically what it does and what it doesn't. Often it is better to say Machine Learning, Natural Language Understanding, Forecasting, Image Recognition, Predictive Analytics/Big Data, etc. I would put these technologies under the banner of Cybernetics, with its transdisciplinary systems and top-down perspective [35].

Having said that, I think it is wise to maintain the dialogue such as ongoing at the Asilomar Conference and World Economic Forum [12], [13], to prepare the way for ethical and safe AGI. Once a self-reinforcing cybernetic loop starts to generate more and more intelligent systems, the onset of AGI is likely to be fast, and catch many unprepared. Given that it is wise to

acknowledge that it is necessary to be prepared to think the unthinkable, before the unstoppable starts, that is if you have a desire to stop it/make it safe.

4.2. Really Useful Machine Learning – RUMLSM

With the rising optimistic and pessimistic media attention to AI, the philosophical debate over the term, and the conflation of AI projects with lofty AGI ambitions, it is perhaps wise to consider using terms that limit reference to AI and AGI at this time. More emphasis should be made on terms such as Machine Learning, Neural Networks, Evolutionary Computation, Vision, Robotics, Expert Systems, Speech Processing, Planning, Natural Language Processing, etc., and to make sure that the scope is well defined in practical application.

In this spirit, we are currently undertaking modest AI research and development of an innovative Hybrid Machine Learning paradigm that incorporates bottom-up Deep Learning Neural Networks and a means to extract a rationalisation of a top-down heuristic narrative. Our intention is to roll this out over 4 phases over the following three years and apply it to calvIO’s Industrial Robotics platforms and integrated systems. We refer, and headline, this as “Really Useful Machine Learning” (RUMLSMSM).

The key novelty of the system is the integration of intuitive bottom-up and rational top-down learning; we find inspiration for the paradigm in known means to teach expertise at an expert practitioner level by “masterclass mentoring”, outlined in the Figure below [36].

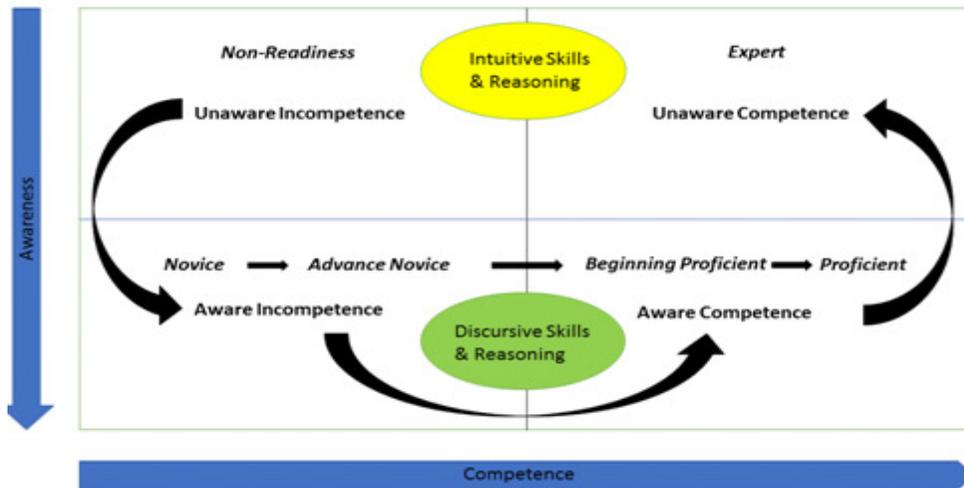


Figure 3: Masterclass Mentoring: Journey from Non-Readiness to Expert

4.3. Compliance, Supervision and Control needs

Key aspects of the deliverable features of RUMLSM will be the ability to manage compliance, supervision and control of the system, by inspection of its extracted rationalised heuristic rule base. The testing of designed systems is performed in extensive simulation scenarios, examining extreme conditions before deploying to the physical systems.

4.4. Extraction of Heuristics from Deep Learning Neural Networks

The means by which Expert Heuristics are extracted from the Deep Learning Neural Networks has been studied by other teams [37], [38] and [39]. The means by which we propose to do so in RUMLSM is an innovative patent pending process. Expert Heuristic/Rule extraction can be defined as "...given a trained neural network and the data on which it was trained, produce a description of the network's hypothesis that is comprehensible yet closely approximates the network's predictive behaviour." Such extraction algorithms are useful for experts to verify and cross-check neural network systems.

Earlier this year, John Launchbury, director of DARPA's Information Innovation Office said, "There's been a lot of hype and bluster about AI." They published their view of AI into "Three Waves", so as to explain what AI can do, what AI can't do, and where AI is headed. RUMLSM is very much in this third wave in our opinion [40].

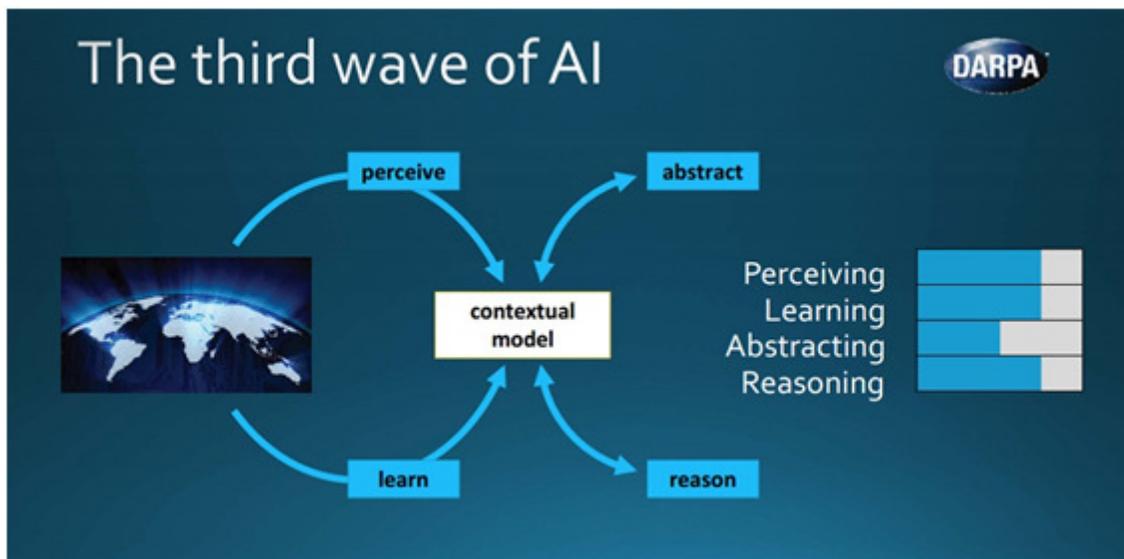


Figure 4: DARPA's Third Wave of AI [Source: 40]

5. CONCLUSIONS

The increase in media interest in Artificial Intelligence (AI) was noted, and the optimistic and pessimistic expectations as to the impact of AI (and potentially future AGI) commented on. While there have been some failures recently, significant milestone successes have encouraged massive investment. But, there is nevertheless a potential for AI to be marked as a bubble and "hyped" technology. Indeed there are already signs of a backlash against AI [41], [42] and [43].

A short-term recommendation, to avoid the extremes of the positive and negative positions, was offered, we should ensure we don't conflate AGI with AI and be more specific in the use of terms to avoid AI becoming "meaningless". As Jerry Kaplan says "Had artificial intelligence been named something less spooky, we'd probably worry about it less." [23]. As an example of a more realistic use of terms for AI, goals is the use of terms such as "Really Useful Machine Learning" (RUMLSM). This particular technique was introduced in the context of resolving

some of the known issues in bottom-up Deep Learning by Neural Networks with a top-down Cybernetic simulation-based process providing a more open and rational means to supervise, regulate and maintain compliance of the system. This patent pending technology uses a Masterclass-Mentoring paradigm, and fits into DARPA's "Third Wave of AI."

ACKNOWLEDGEMENTS

The author would like to thank Jeff McCormack CEO and colleagues Nicole Tedesco and Alex Stickle of calvIO Inc for their support and input into this paper.

REFERENCES

Note: Given the media content referenced the format has been adjusted to feature the links to these.

- [1] Ethan Fast, Feb 3 2017, "Long-term Trends in the Public Perception of Artificial Intelligence", <https://hackernoon.com/long-term-trends-in-the-public-perception-of-artificial-intelligence-6b1512fdd7ac#.v16cpe36l> Last Referenced 5th March 2017.
- [2] Fortune Tech, Jun 03, 2016, "Tech CEOs Declare This the Era of Artificial Intelligence" <http://fortune.com/2016/06/03/tech-ceos-artificial-intelligence/> Last Referenced 5th March 2017.
- [3] Gil Press, 11 November 2016, "Forrester Predicts Investment In Artificial Intelligence Will Grow 300% in 2017" <https://www.forbes.com/sites/gilpress/2016/11/01/forrester-predicts-investment-in-artificial-intelligence-will-grow-300-in-2017/#44642b785509> Last Referenced 5th March 2017.
- [4] Om Malik, 26 August 2016, "The Hype—and Hope—of Artificial Intelligence", <http://www.newyorker.com/business/currency/the-hype-and-hope-of-artificial-intelligence> Last Referenced 5th March 2017.
- [5] H. B. Barlow, 21 July 1983, "Intelligence, guesswork, language", *Nature* 304, 207 - 209 (21 July 1983); doi:10.1038/304207a0, (also see... <http://www.nature.com/nature/journal/v304/n5923/abs/304207a0.html> Last Referenced 5th March 2017.)
- [6] Fatmi, H.A. & Young, R.W. *Nature* 228, 97 (1970) (also see... <http://www.ifsr.org/index.php/in-memorandum-dr-haneef-akhtar-fatmi-03-july-1933-04-april-1995/>)
- [7] Richard Mallah, 29 December 2015, "The Top A.I. Breakthroughs of 2015", <https://futureoflife.org/2015/12/29/the-top-a-i-breakthroughs-of-2015/> Last Referenced 5th March 2017.
- [8] Mona Lalwani, 25 December 2016 "AI was everywhere in 2016" <https://www.engadget.com/2016/12/25/ai-was-everywhere-in-2016/> Last Referenced 5th March 2017.
- [9] Tonya Riley, 3 March 2017, "Artificial intelligence goes deep to beat humans at poker" <http://www.sciencemag.org/news/2017/03/artificial-intelligence-goes-deep-beat-humans-poker> Last Referenced 5th March 2017.
- [10] 5th December 2015 "MIRI AI Predictions Dataset" <http://aiimpacts.org/miri-ai-predictions-dataset/> Last Referenced 5th March 2017.
- [11] Peter Diamandis, Steven Kotler, 21 February 2012, ISBN: 9781451614213, Publishers: Free Press, Tantor Media "Abundance: The Future Is Better Than You Think", See... <http://www.diamandis.com/abundance/> Last Referenced 5th March 2017.
- [12] Future of Life Institute, "BENEFICIAL AI 2017" Asilomar Conference 2017 <https://futureoflife.org/bai-2017/> Last Referenced 5th March 2017.
- [13] The Future of Artificial Intelligence and Robotics, World Economic Forum 2017, <https://www.weforum.org/communities/the-future-of-artificial-intelligence-and-robotics> Last Referenced 5th March 2017.
- [14] A Jung Moon, "RobotEthics info Database" <http://www.amoon.ca/Roboethics/> Last Referenced 5th March 2017.

- [15] Alex Hern, 17 January 2017, "Give robots 'personhood' status, EU committee argues", <https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues> Last Referenced 5th March 2017.
- [16] John West, 2 April 2016, "Microsoft's disastrous Tay experiment shows the hidden dangers of AI", <https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/> Last Referenced 5th March 2017.
- [17] Andrew Orlowski 22 Feb 2017, "Facebook scales back AI flagship after chatbots hit 70% f-AI-lure rate", https://www.theregister.co.uk/2017/02/22/facebook_ai_fail/ Last Referenced 5th March 2017.
- [18] Justin Bariso, 23 February 2017, "Microsoft's CEO Sent an Extraordinary Email to Employees After They Committed an Epic Fail", <http://www.inc.com/justin-bariso/microsofts-ceo-sent-an-extraordinary-email-to-employees-after-they-committed-an-.html> Last Referenced 5th March 2017.
- [19] James Vincent, 10 October 2016, "These are three of the biggest problems facing today's AI", <http://www.theverge.com/2016/10/10/13224930/ai-deep-learning-limitations-drawbacks> Last Referenced 5th March 2017.
- [20] IEEE Technical Community Spotlight, 2 September 2016, "On the Use of AI – the Dependency Dilemma", <http://sites.ieee.org/spotlight/ai-ethical-dilemma/> Last Referenced 5th March 2017.
- [21] Josh Weisberg, Internet Encyclopaedia of Philosophy, "The Hard Problem of Consciousness", <http://www.iep.utm.edu/hard-con/> Last Referenced 5th March 2017.
- [22] Daniel Dennett, 28 February 2017, "Why robots won't rule the world – Viewsnight", BBC Newsnight, <https://www.youtube.com/watch?v=2ZxzNAEFtOE>, Last Referenced 5th March 2017.
- [23] Jerry Kaplan, 3 March 2017, AI's PR Problem, <https://www.technologyreview.com/s/603761/ais-pr-problem/> Last Referenced 5th March 2017.
- [24] Tamer Sameeh, 6 March 2017, "Decentralized Artificial Super-intelligence Via Multi-agent Systems and Ethereum's Smart Contracts" <https://steemit.com/crypto-news/@tamersameeh/decentralized-artificial-super-intelligence-via-multi-agent-systems-and-ethereum-s-smart-contracts> Last Referenced 6th March 2017.
- [25] Luzius Meisser, 11 January 2017, "Simulations as Test Environments" <http://meissereconomics.com/2017/01/11/Simulation.html> Last Referenced 6th March 2017.
- [26] Tom Simonite, 13 May 2016, MIT Technology Review, "Moore's Law Is Dead. Now What?" <https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/> Last Referenced 6th March 2017.
- [27] Justine Brown, 6 January 2017, CIO Dive, "Moore's Law is 'alive and well and flourishing,' Intel CEO says" <http://www.ciodive.com/news/moores-law-is-alive-and-well-and-flourishing-intel-ceo-says/433484/> Last Referenced 6th March 2017.
- [28] Josh Hamilton, February 13, 2017 "Moore's Law and Quantum Computing" <https://cloudtweaks.com/2017/02/moores-law-quantum-computing/> Last Referenced 6th March 2017.
- [29] Michael Sainato, 19 August 2015 Observer, "Stephen Hawking, Elon Musk, and Bill Gates Warn About Artificial Intelligence" <http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/> Last Referenced 6th March 2017.
- [30] Rory Cellan-Jones, 2 December 2014, BBC, "Stephen Hawking warns artificial intelligence could end mankind" <http://www.bbc.co.uk/news/technology-30290540> Last Referenced 6th March 2017.
- [31] George Dvorsky, 14 January 2015, "Prominent Scientists Sign Letter of Warning About AI Risks" <http://io9.gizmodo.com/prominent-scientists-sign-letter-of-warning-about-ai-ri-1679487924> Last Referenced 6th March 2017.
- [32] Gil Press, 26 February 2017, Forbes, "Luddites Against Job-Killing Automation And Technology Enthusiasts Creating New Industries" <https://www.forbes.com/sites/gilpress/2017/02/26/luddites-against-job-killing-automation-and-technology-enthusiasts-creating-new-industries/#425e0ea77e46> Last Referenced 6th March 2017.
- [33] Tom Peters, 1987 "Under Promise, Over Deliver" <http://tompeters.com/columns/under-promise-over-deliver/> Last Referenced 6th March 2017.
- [34] Ian Bogost 4 March 2017, The Atlantic, "'Artificial Intelligence' Has Become Meaningless" <https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547/> Last Referenced 6th March 2017.

- [35] Kevin Warwick, 10 November 2016, MIT Technology Review, “The Future of Artificial Intelligence and Cybernetics” <https://www.technologyreview.com/s/602830/the-future-of-artificial-intelligence-and-cybernetics/> Last Referenced 6th March 2017.
- [36] D’Youville College, Professional Development, <https://www.dyc.edu/academics/professional-development/> Last Referenced 6th March 2017.
- [37] Tameru Hailesilassie, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 14, No. 7, July 2016 “Rule Extraction Algorithm for Deep Neural Networks: A Review” <https://arxiv.org/ftp/arxiv/papers/1610/1610.05267.pdf> Last Referenced 6th March 2017.
- [38] Jan Ruben Zilke, Master Thesis, TUD, “Extracting Rules from Deep Neural Networks” http://www.ke.tu-darmstadt.de/lehre/arbeiten/master/2015/Zilke_Jan.pdf Last Referenced 6th March 2017.
- [39] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, Eric P. Xing, School of Computer Science Carnegie Mellon University, 2016 “Harnessing Deep Neural Networks with Logic Rules” http://www.cs.cmu.edu/~epxing/papers/2016/Hu_etal_ACL16.pdf Last Referenced 6th March 2017.
- [40] Steve Crowe, 21 February 21 2017, Robotics Trends, “What AI Can and Can’t Do: DARPA’s Realistic View,” http://www.roboticstrends.com/article/what_ai_can_and_cant_do_darpas_realistic_view/Artificial_Intelligence
- [41] Matt Asay., 3 March 2017, Infoworld Tech Watch, “Artificially inflated: It’s time to call BS on AI” <http://www.infoworld.com/article/3176602/artificial-intelligence/artificially-inflated-its-time-to-call-bs-on-ai.html> Last Referenced 6th March 2017.
- [42] Andrew Orlowski, 2 January 2017, The Register, “‘Artificial Intelligence’ was 2016’s fake news” https://www.theregister.co.uk/2017/01/02/ai_was_the_fake_news_of_2016/ Last Referenced 6th March 2017.
- [43] Luke Dormehl, 18 February 2017, Wired, “Don’t believe the hype when it comes to AI” <http://www.wired.co.uk/article/sensationalism-ai-hype-innovation> Last Referenced 6th March 2017.

ANKLE MUSCLE SYNERGIES FOR SMOOTH PEDAL OPERATION UNDER VARIOUS LOWER-LIMB POSTURE

Kazuo Kiguchi¹, Takuto Fujita¹, Sho Yabunaka², Yusaku Takeda²,
and Toshihiro Hara²

¹Department of Mechanical Engineering, Kyushu University, Fukuoka, Japan

²Mazda Motor Corporation, Fuchu-cho, Hiroshima, Japan

ABSTRACT

A study on muscle synergy of ankle joint motion is important since the acceleration operation results in automobile acceleration. It is necessary to understand the characteristics of ankle muscle synergies to define the appropriate specification of pedals, especially for the accelerator pedal. Although the biarticular muscle (i.e., gastrocnemius) plays an important role for the ankle joint motion, it is not well understood yet. In this paper, the effect of knee joint angle and the role of biarticular muscle for pedal operation are investigated. Experiments of the pedal operation were performed to evaluate the muscle synergies for the ankle plantar flexion motion (i.e., the pedal operation motion) in the driving position. The experimental results suggest that the muscle activity level of gastrocnemius varies with respect the knee joint angle, and smooth pedal operation is realized by the appropriate muscle synergies.

KEYWORDS

Ankle Joint, Pedal Operation, EMG, BiArticular Muscle, Muscle Synergy

1. INTRODUCTION

In order to design an automobile, human characteristics such as human musculoskeletal characteristics must be taken into account. Control of ankle joint motion is important for many tasks such as operation of an automobile. In the case of automobile operation, acceleration of the automobile is controlled with an acceleration pedal using the driver's ankle joint motion. Therefore, it is important to understand the characteristics of ankle muscle synergies to define the appropriate specification of pedals, especially for the accelerator pedal since its operation directly results in the acceleration of the automobile. Although the ankle joint motion is generated with several muscles, it is known that the biarticular muscle such as the gastrocnemius plays an important role for the joint motion [1]-[4]. Since gastrocnemius is a biarticular muscle, not only the ankle joint angle, but also the knee joint angle affects the muscle activity of gastrocnemius [4]. Therefore, lower-limb posture must be taken into account to consider the role of gastrocnemius muscle activity for the ankle joint motion, especially for plantar flexion motion. Consequently, the driving position affects the performance of muscle synergy of the ankle joint. However, the role of biarticular muscles is still not well understood [5] although the role of gastrocnemius activity for the ankle joint motion must be considered to understand ankle muscle synergies, especially for accelerator pedal operation.

In this paper, the effect of knee joint angle and the role of biarticular muscle for pedal operation are investigated. Experiments of the pedal operation were performed to evaluate the muscle synergies for the ankle plantar flexion motion (i.e., the pedal operation motion) in the driving position. Electromyographic (EMG) activities of the certain important muscles were measured under several lower-limb postures (i.e., under several driving position) considering the pedal operation of an automobile in the experiment. The experimental results suggest that the muscle activity level of gastrocnemius varies with respect the knee joint angle, and smooth pedal operation is realized by the appropriate muscle synergies. This study is also important to control robotic devices [6] to assist the ankle joint motion according to the user's motion intention since EMG activities reflect the user's motion intention [7].

2. EMG ACTIVITIES FOR PEDAL OPERATION

Experiments were performed to measure EMG activities of the certain important muscles of the ankle joint during the pedal operation motion in the driving position. In the experiment, the human subjects sat on the driving seat with the several different driving positions and then performed the pedal operation to investigate the effect of lower-limb posture. The relation between the accelerator pedal operation motion and the EMG activities of the certain muscles of the ankle joint was measured in the experiment.

2.1. Experimental Setup

The experimental setup is shown in Fig. 1. It mainly consists of a driving seat, an accelerator pedal, and a display. The relative position between the driving seat and the accelerator pedal can be changed. The ankle joint angle is measured with a goniometer. In the experiment, an ordinal accelerator pedal is used and the scenery on the display is changed in accordance with the pedal operation by the driver.

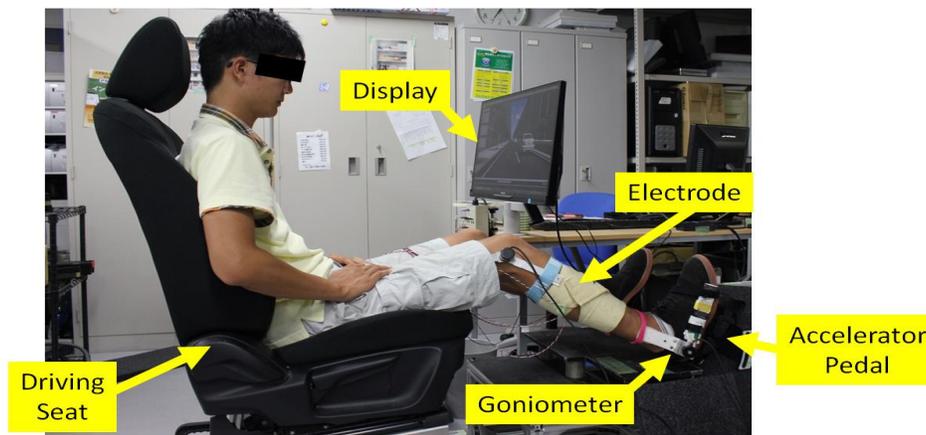


Figure 1. Experimental setup

2.2. Measurement

The EMG signals of the important muscles for the ankle joint motion (i.e., tibialis anterior, peroneus longus, gastrocnemius, and soleus) are measured in the experiment. The location of each electrode for EMG measurement is shown in Fig. 2. The EMG signals from the electrodes are amplified and sent to the computer. In order to extract a feature of the EMG signal, the Root Mean Square (RMS) is calculated as shown below.

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N v_i^2} \quad (1)$$

where N is the number of samples in a segment, v_i is the voltage at i^{th} sampling point. The number of samples N is set to be 100 and the sampling frequency is 500Hz in this study. Then it is transferred to %MVC.

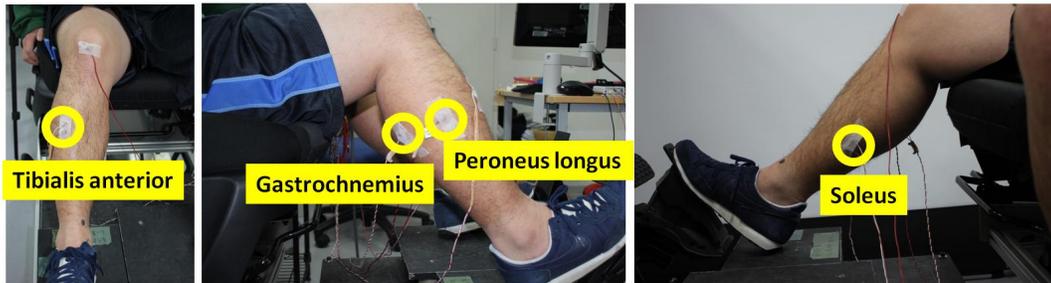


Figure 2. Location of each electrode

2.3. Experiment of Pedal Operation

Experiments were carried out with three human subjects. Basically, three types of seat position (sedan, SUV, and sport types) are prepared for the experiments considering the types of automobile as shown in Fig. 3. The angle of the acceleration pedal is also different in each seat type. Another three kinds of seat slide position (close, appropriate, and far positions) are prepared in the experiments considering the driving position in an automobile. Examples of these positions in the sedan type are shown in Fig. 4. The details of the initial lower-limb posture are written in Table 1. Here, the knee angle means the knee joint flexion angle from the extended knee position and the ankle angle means the ankle joint plantar flexion angle with respect to the shank axis.

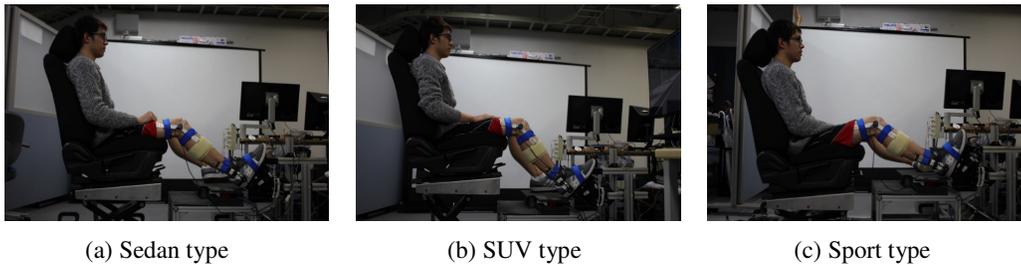


Figure 3. Seat Positions

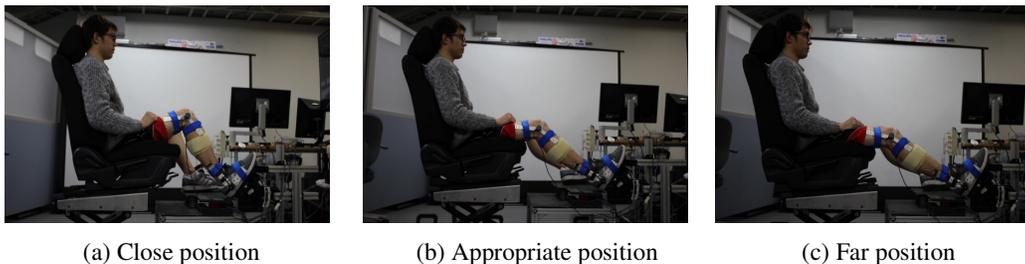


Figure 4. Seat Slide Positions (Driving Positions) in Sedan Type

In the experiment, another automobile is running at 60 [km/h] to the same direction in the next lane on the display. The subject (driver) operates the accelerator pedal to accelerate the automobile from the state of rest to catch up with another automobile in the next lane. Then, the subject operates the accelerator pedal to make the automobile run side by side in the first experiment.

In the second experiment, the subject (driver) catches up with another automobile which is running at 60 [km/h] in the next lane and makes the automobile run side by side for a while, then overtakes another automobile in the next lane after a while.

Table 1. Initial Lower-Limb Postures.

Subject (Height)	Seat Type	Seat Position	Knee Angle [deg]	Ankle Angle [deg]
A (1.79 m)	Sedan	Close	91.7	68.4
		Appropriate	137.5	80.8
		Far	149.6	87.8
	SUV	Close	99.2	61.3
		Appropriate	133.8	82.9
		Far	146.5	87.6
	Sport	Close	100.9	76.6
		Appropriate	146.9	90.9
		Far	155.9	93.7
B (1.66 m)	Sedan	Close	98.7	70.8
		Appropriate	127.6	85.2
		Far	145.3	89.9
	SUV	Close	99.2	61.3
		Appropriate	122.7	82.9
		Far	140.1	87.6
	Sport	Close	109.6	80.2
		Appropriate	131.4	96.1
		Far	149.7	113.4
C (1.78m)	Sedan	Close	100.6	74.1
		Appropriate	120.2	86.8
		Far	150.7	96.9
	SUV	Close	98.5	66.1
		Appropriate	119.4	82.8
		Far	137.5	85.9
	Sport	Close	100.9	78.2
		Appropriate	115.4	92.6
		Far	159.7	100.5

3. EXPERIMENTAL RESULTS

Figures 5-7 show the results of the first experimental of the subject A with the sedan, SUV, and sport type seat positions, respectively. These results show that the activities of muscles for plantar flexion such as peroneus longus, soleus, and gastrocnemius increase as the seat slide position becomes further from the acceleration pedal since the knee joint is extended and the ankle dorsiflexion angle is decreased. Especially, the increase the ratio of monoarticular muscle peroneus lingus activity is prominent. The ratio of biarticular muscle gastrocnemius varies depends on the seat position. These results show that the muscle synergy of the ankle joint motion for the same acceleration pedal operation varies in accordance with the knee angle and the ankle angle of the driver.

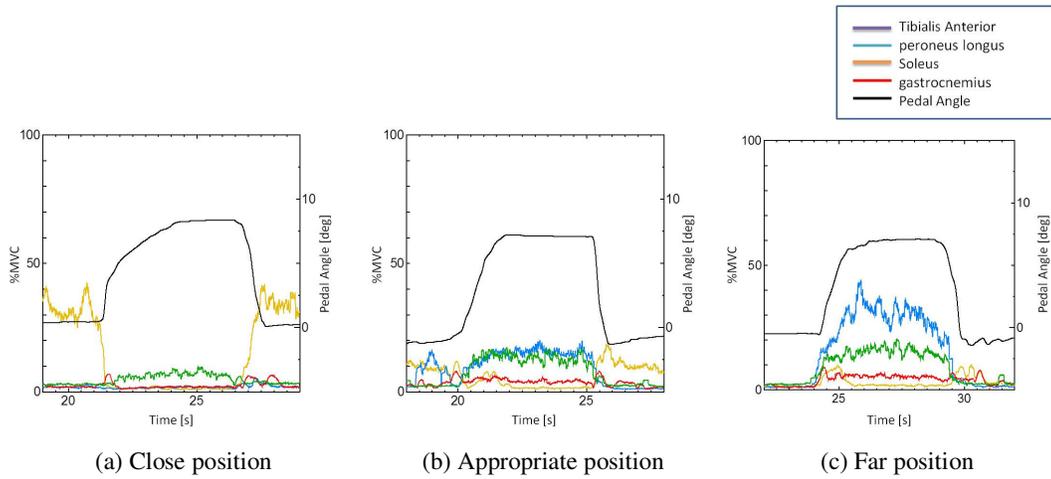


Figure 5. Experimental Results in Sedan Type (Subject A)

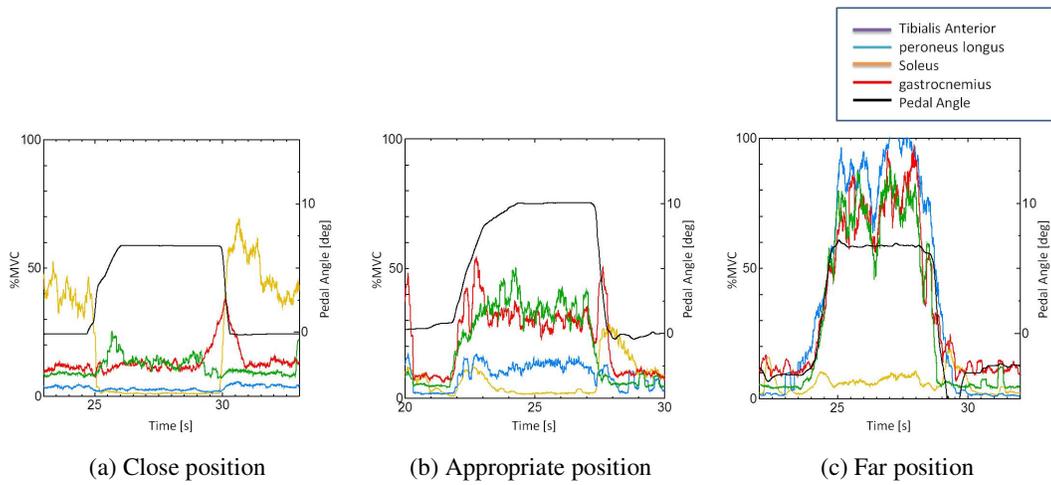


Figure 6. Experimental Results in SUV Type (Subject A)

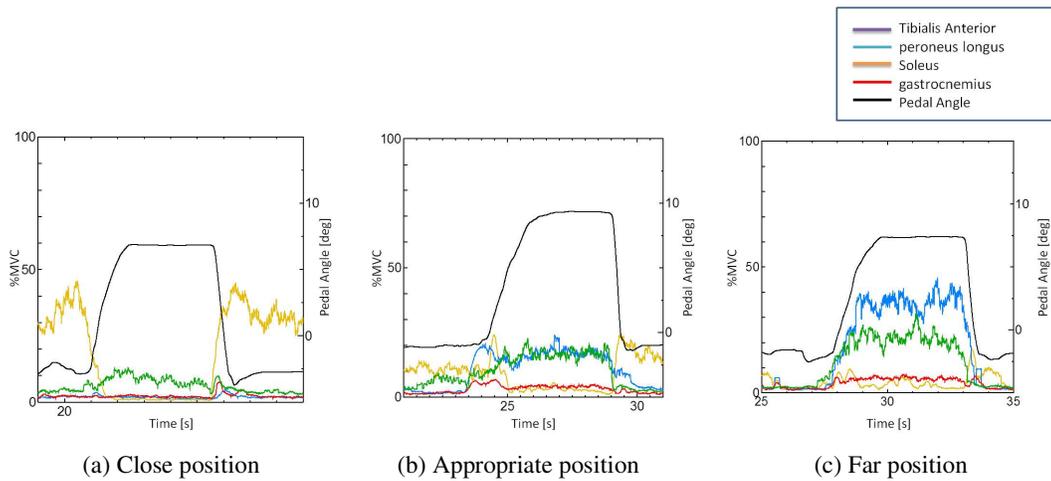


Figure 7. Experimental Results in Sport Type (Subject A)

The angle of the acceleration pedal also affects the muscle synergy of the ankle joint. Even though the initial lower-limb posture in the far seat position in the SUV seat type and that in the appropriate seat position in the sport seat type of the subject A is almost the same, the muscle activity level is different since the angle of the acceleration pedal is different. The same tendency can be observed with the experimental results with the other subjects.

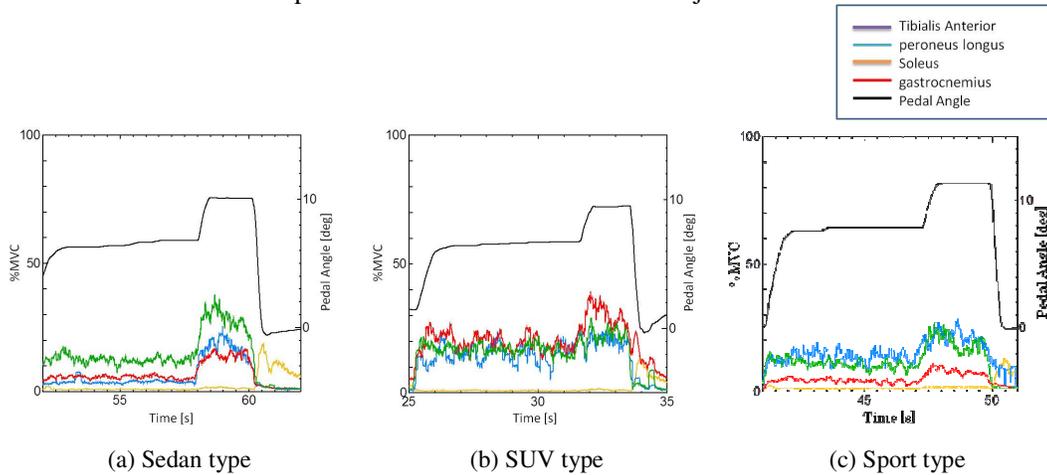


Figure 8. Experimental Results of Overtaking (Subject A)

The results of the second experimental of the subject A in the appropriate seat position with the sedan, SUV, and sport type seat are shown in Fig. 8. The results show that the activity levels of muscles for plantar flexion such as peroneus longus, soleus, and gastrocnemius increase when the angle of the acceleration pedal is increased for the overtaking. Note that the ratio of each muscle for the plantar flexion varies when the angle of the acceleration pedal is increased for the overtaking.

4. DISCUSSION

The experimental results show that the activity levels of muscles for plantar flexion such as peroneus longus, soleus, and gastrocnemius increase as the seat slide position becomes further from the acceleration pedal since the knee joint is extended and the ankle dorsiflexion angle is decreased. As one can see in Table 1, the most of the initial ankle joint angles are slightly dorsiflexed angles, especially in the case of the close seat position. Therefore, the muscle for the dorsiflexion (i.e., tibialis anterior) is released to make the plantar flexion motion. In this case, the plantar flexion motion can be generated with a little activity of the muscles for plantar flexion. Since the gastrocnemius is a biarticular muscle, the knee joint angle affects the activity level. Arampatzis *et. al* [4] also suggest that a critical force-length potential of muscles of the triceps surae results in the decrease of EMG activity of the gastrocnemius medialis at pronounced knee flexed positions.

As the seat slide position becomes further from the acceleration pedal, the initial ankle joint angles become closer to the plantar flexed angles. Therefore, further plantar flexion motion is necessary to operate the acceleration pedal. Consequently, the muscle activity levels of the muscles for the plantar flexion become higher. The experimental results shown in Figs. 5-7 suggest that the ratio of each muscle for the plantar flexion effectively varies in accordance with the lower-limb posture and the angle of the accelerator pedal to make the smooth pedal operation. The experimental results shown in Fig. 8 suggest that the ratio of each muscle for the plantar flexion moderately varies when the angle of the acceleration pedal is increased for the overtaking

since the reaction force from the pedal is increased. When the angle of the acceleration pedal is increased, the activity level of the gastrocnemius becomes a little higher instantly sometimes. Ingen Schenau *et. al* [1] showed that the biarticular muscles such as the gastrocnemius are used to transport energy from proximal to distal joints during jumping. Therefore, the change of the muscle synergy might be concerned with the energy transmission. The experimental results suggest that the muscle synergy of the ankle plantar flexion motion for the acceleration pedal operation is moderately controlled according to the condition. Further study is required to understand the muscle synergy of ankle joint motion for acceleration pedal operation

5. CONCLUSIONS

In this paper, the effect of knee joint angle and the role of biarticular muscle for ankle joint motion (i.e., pedal operation) were investigated. The experimental results showed that muscle synergy of ankle joint motion for acceleration pedal operation moderately varies in accordance with the condition such as the lower-limb posture of the driver, the angle of the acceleration pedal, and the amount of the required ankle torque.

REFERENCES

- [1] Ingen Schenau, G.J. van, Bobbert, M.F., & Rozendal, R.H., (1987) “The unique action of bi-articular muscles in complex movements”, *Journal of Anatomy*, vol. 155, pp1-5.
- [2] Kumamoto, M., Oshima, T., & Yamamoto, T., (1994) “Control properties induced by the existence of antagonistic pairs of bi-articular muscles – Mechanical engineering model analyses”, *Human Movement Science*, vol. 13, pp611-634.
- [3] Lee, S.S.M. & Piazza, S.J., (2008) “Inversion–eversion moment arms of gastrocnemius and tibialis anterior measured in vivo”, *Journal of Biomechanics*, vol. 41, pp3366-3370.
- [4] Arampatzis, A., Karamanidis, K., Stafilidis, S., Morey-Klapsing, G., DeMonte, G., & Bruggemann, G.P., (2006) “Effect of different ankle- and knee-joint positions on gastrocnemius medialis fascicle length and EMG activity during isometric plantar flexion”, *Journal of Biomechanics*, vol. 39, pp1891-1902.
- [5] Cleather, D.J., Southgate, D.F.L., Stafilidis, S., & Bull, A.M.J., (2015) “The role of the biarticular hamstrings and gastrocnemius muscles in closed chain lower limb extension”, *Journal of Theoretical Biology*, vol. 365, pp217-225.
- [6] Jimenez-Fabian, R. & Verlinden, O., (2012) “Review of control algorithms for robotic ankle systems in lower-limb orthoses, prostheses, and exoskeletons”, *Medical Engineering & Physics*, vol. 34, pp397-408.
- [7] Kiguchi, K. & Hayashi, O., (2012) “An EMG-Based Control for an Upper-Limb Power-Assist Exoskeleton Robot”, *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, Vol. 42, no. 4, pp1064-1071

AUTHORS

Kazuo Kiguchi received the B.E. degree in mechanical eng. from Niigata Univ., Japan in 1986, the M.A.Sc. degree in mechanical eng. from the Univ. of Ottawa, Canada in 1993, and the Doctor of .Eng. degree in Mechano-Informatics from Nagoya Univ., Japan in 1997. He is currently a professor in the Dept. of Mechanical Engineering, Faculty of Eng., Kyushu University, Japan. His research interests include biorobotics, human assist robots, and health care robots.



Takuto Fujita is currently a student in mechanical and aerospace eng., Faculty of Eng., Kyushu Univ., Japan



Sho Yabunaka received the BSc. degree in Culture and Information Science from Doshisha University Japan, in 2010, and MBIT. degree in Information Science from Nara Institute of Science and Technology, Japan, in 2012. He is currently a Specialist of Kansei Engineering Research, Advanced Human-Vehicle Research Field, Technical Research Center, Mazda Motor Corporation. His current research interests include the mechanism of human musculoskeletal system and these applications for designing vehicle.



Yusaku Takeda received the B.E. and M.E degrees in mechanical systems and design eng. from Hiroshima Univ., Japan, in 2000 and 2003, and Doctor of Engineering degree in Artificial Complex Systems Eng. from Hiroshima Univ. in 2005. He is currently a Senior Specialist of Kansei Engineering Research, Advanced Human-Vehicle Research Field, Technical Research Center, Mazda Motor Corporation. His current research interests include the mechanism of human musculoskeletal system and visual awareness, and these applications for designing vehicle.



Toshihiro Hara received the B.E. degrees in electrical engineering from Himeji Institute of Technology, Japan, in 1991. He is currently a Technical Leader of Kansei Engineering Research, Advanced Human-Vehicle Research Field, Technical Research Center, Mazda Motor Corporation. His current research interests include the novel HMI design method based on the ergonomics, and these applications for designing vehicle.



ADVANCED LSB TECHNIQUE FOR AUDIO STENOGRAPHY

Mohammed Salem Atoum¹, Mohammad M Alnabhan² and Ahmad
Habboush³

¹Faculty of Science, Information Technology and Nursing,
Irbid National University, Irbid, Jordan

²Faculty of Information Technology, Mutah University

³Faculty of Information Technology, Jerash University

ABSTRACT

This work contributes to the multimedia security fields by given that more protected steganography technique which ensures message confidentiality and integrity. An Advanced Least Significant Bit (ALSB) technique is presented in order to meet audio steganography requirements, which are imperceptibility, capacity, and robustness. An extensive evaluation study was conducted measuring the performance of proposed NLSB algorithm. A set of factors were measured and used during evaluation, this includes; Peak Signal to Noise Ratio (PSNR) and Bit Error Rate. MP3 Audio files from five different audio generators were used during evaluation. Results indicated that ALSB outperforms standard Least Significant Bit (SLSB) technique. Moreover, ALSB can be embedding an utmost of 750 kb into MP3 file size less than 2 MB with 30db average achieving enhanced capacity capability.

KEYWORDS

MP3, LSB, Cryptography, Steganography.

1. INTRODUCTION

There is a continuous challenge in securing digital transmission between network nodes against any form of penetration and intrusion. Ensuring security of information requires considering three main components confidentially, integrity, and availability. This can be conducted through techniques, described as steganography and cryptography [1].

Cryptography is the process of encryption and decryption of a digital data. However, cryptography techniques are considered weak or consume high resources. Steganography is mainly based on covering digital data in a safe digital carrier [2]. Steganography is utilized for hiding secret messages in ancient times [3]. According to [4], steganography can be described as a method of hiding secondary information (e.g. file or message) within primary information, known as the carrier or host, with no effect on the size of information and without causing any form of distortion. The information is embedded within a media expressed as a bit stream, stego signal or sequence [5].

Watermarking is another technique that is used to insert watermark into host cover to protect information such as copyright for hosts [6]. Steganography and watermarking usually embed information in host media in a transparent manner [7]. However, considering watermarking, the process requires compromising intentional attacks and preventing any cause of information destruction by insuring robustness and protecting signals quality [6]. Watermarking is the most suitable technique in scenarios where hidden information knowledge can result in information manipulations [7].

The strength of steganographic technique is based on saving the data in the carrier medium against attacks or alteration. Audio files are considered very suitable media providing different compression rates and allow performing steganography in MP3 format. Audio stenographic methods based on SLSB have gained continuous concern. However, this technique has limitations in security, capacity and imperceptibility. In addition, to date, embedding messages after audio compression has not been widely considered. Accordingly, this work investigates standard audio steganographic techniques and addresses its weaknesses and presents an Advanced Least Significant Bit (ALSB) technique in order to improve robustness and security of the standard LSB algorithm.

2. STANDARD LEAST SIGNIFICANT BIT

Standard Least significant bit (SLSB) algorithm is considered simplest steganographic method [8]. In SLSB, secret message and cover are converted to stream of bits. One or more bit of secret message are used to replace cover LSB bits. Afterwards, bits stream are sent to the receiver which uses the same algorithm to extract the embeded message [9]. SLSB uses least significant bits of the cover and utilizes sequential embedding. This result in clear suspicion secret message location within the cover files [10]. Hence, it is easier to detect and extract the secret message by the attacker [11]. In order to enhance the efficiency of SLSB algorithm security, a generator described as pseudorandom number (PN) is used [12]. However, the use of PN has incurred time limitations, because using PN requires more time to operate.

A few research works have been conducted in the area of MP3 audio steganography more specially while considering embedding after compression [11]. The cause might be the weakness of this technique in achieving a good well expansion of information data steganography, and in some cases results in low quality sound. The MP3 file is compression file that means is not flexible as well as the size is less compared to other audio file types [12]. The Embedding secret message by using after compression methods is able to create audio corruption. Two methods after compression are used to embed secret message: embedding in header frames and embedding in audio data.

2.1 EMBEDDING IN HEADER FRAMES

Before describing methods using header frames to embed secret message, MP3 file structure is explained. MP3 file is divided into header frames and audio data frames. Techniques used for encoding MP3 files are: constant bit rate CBR, average bit rate ABR and variable bit rate VBR. These methods are expected to use padding bytes. Several methods have utilized unused bits in header frames and padding bytes before all frames and between frames, in order to replace bits from secret message. However, weaknesses of these methods include; limited capacity and

security. Using padding stuffing byte technique [13], the sender converts empty information in padding byte in the cover with secret message. However, the capacity of embedded secret message depends on the size of padding byte, which was added in the cover file using encoding methods: CBR, VBR and ABR. At the receiver side, information search within stego file is applied to find the location of padding byte in the cover and extract the secret message. Unused bits in header frames such as private bit, copyright bit, original bit, and emphasis bit can be used to embed secret message without affects the quality of sound [13]. This technique replaces bit from secret message with a bit in the header. However, using this technique it is easily to extract the secret message from the header and change it from attackers.

Using Before All Frames (BAF) technique [14], researchers develop new technique to embed hole secret message before the first frames in the header. The secret message with encrypted text is embedded in a cover file, will have a maximum size of 15 KB, however the size will reach 30 KB without using encryption. This technique is better capacity compared with padding and unused bit, but also is less security without using encryption method before embedding the secret message. In addition, Between Frames (BF) methods divide the secret message before embedding it into small size cover file [14]. This method depends on the size of secret message, and on the spaces between frames of the cover file. The maximum size of secret message can be embedded is not fixed, because it can expanded the size of the cover file. The advantages of BF technique are high capacity and imperceptibility, but the disadvantages are less security and robustness. It can be concluded that all methods of header frame are facing limited robustness against attackers [14].

2.2 EMBEDDING IN AUDIO DATA:

Several methods have addressed security problems in embedding the secret message in audio data using header frames. [14] presents a new method that embeds one, two, three or four bits from MP3 file by replacing one or two or three or four bits from the secret message, described in text format. The first byte from the cover file is selected randomly. Using this method, random position in the cover file is chosen to start embedding the secret message. This is sequentially repeated to embed the secret message in the cover file. The drawback for using this method is limited robustness as well as the random position it was used is not permanent with a fixed size.

3. PROPOSED ALGORITHM

To address limitations of embedding algorithms after compression, this work introduces a new technique in LSB. The algorithm is described as Advanced Least Significant Bit (ALSB) technique, which is developed to increase the security of secret message, and improves the method of embedding the secret message in the host file. The main problem in LSB is its weaknesses against intentional or unintentional attacks on the secret message. In ALSB, the sender side uses random position selection from initial 100 byte from the host file. Moreover, the value of LSB and MSB is used to select the bit location required to be embedded in the secret message. If LSB and MSB have the same value, ALSB uses 4 bits from the secret message in order to embed from location two to five of each byte. Otherwise, the technique uses just two bits from secret message to embed it in location two and three. This methodology has increased the security of LSB. The ALSB algorithm pseudo code is discussed below:

Algorithm: Advanced Least Significant Bit Algorithm

```

1: // H is the host file and SM is the secret message and H,SM are the inputs.
2: // H' is the host file (H+SM) and H' is output
3: // beginning to read host file H from initial bit and save it in H'.
4:     start
5:   For i = 1 to Size of (H) do
6:       { H'i ← Hi
7:         }
8: // Create random position from earliest 100 byte in the host file by
   using random generation method Rp
9: // H' is the input
10: // Rp is the output
11:     For i=1 to 100
12:     do
13: // choosing the random byte
14:   { Rp = position ( i )
15:     }
16: // begin to create H' by using ALSB technique to insert message blocks MB
17:     For i= Rp to size of (H')
18:     do
19:   { For j=1 to L (MB)
20:     do
21:   { Read the LSB and MSB value from the byte }
22:   if the LSB+MSB = 00 or 11 then
23:   { embed MB from 2nd to 5th position }
24:   else if LSB+MSB= 10 or 01 then
25:   { embed MB from 3rd to 4th }
26:   Go to next byte
27:   }
28:   }

```

After the sender side implements ALSB technique, the stego object is constructed. To evaluate stego object before send it via internet, the PSNR and BER methods are used to introduce the results of noise in stego object. At the receiver side, inverse method is applied to predict the secret message from the stego object. This prediction is based on the secret information received from safe channel.

4. MEASUREMENT AND EVALUATION

This section describes main measurement metrics used to evaluate the proposed NLSB in terms of reliability, imperceptibility and performance. These metrics include peak signal-to-noise ratio (PSNR), and Bit Error Rates (BER). PSNR is the statistical value (ratio) that compared a signal's maximum power with power of the signal's noise, it logarithmic scale and expressed in decibels (db) [15]. However, PSNR is the peak error measure. The PSNR is error metrics used for quality measurement. The PSNR value of the new algorithm is compared with PSNR of the SLSB algorithm. Low when PSNR value is high, this describes better quality [16]. The PSNR equation is shown below:

$$PSNR = 10 \log_{10} \frac{(\text{MAX}(\text{cov}(i)))^2}{MSE} \quad (1)$$

Where MAX is the maximum differentiation of the inputs (host file sample) compared with stego object in order to validate if the stego object holds secret data or not.

The second metric used is Bit Error Rates (BER) which measures bit errors in a file, beside the summation number of bits required for the spread during a time period. The BER is the statistical measures in telecommunication transmission, the ratio result is percentage of a bit including errors comparing to entire bits. This measure is expressed as ten to a negative power. If the results is low data rate that means is very high in a transmission [15]. In addition, the BER is a measure of bit error probability. BER is considered more accurate while considering increased number of bit errors and long interval. BER is also used to measure the successful recovery of the hidden information. This will have its high effect in real communication channel where errors exists retrieving hidden information. BER is computed as described in the following equation:

$$BER = \frac{1}{Z(\text{cov})} * \sum_{i=0}^{L(H)} (H(i) - H'(i)) \quad (2)$$

Where L is the length, H is host file and H' is stego object.

Table 1 describes audio generators used during the experiment and explains specifications of each audio file including duration in minutes and size in Mbps. These audio clips were used in the evaluation study to measure the effectiveness of the proposed ALSB technique comparing to Standard LSB (SLSB) and XLSB techniques.

Table 1. Audio file specifications

Name of Audio generator	Time (Minute)	Size under 320kbps (MB)
Pop	3:10	8.9
Rock	3:40	9.9
Blues	3:45	10.7
Hip-hop	4:30	12.4
Dance	5:30	14.2
Metal	7:00	14.8

As shown in table 2, proposed ALSB achieved high PSNR values comparing to SLSB [17] and eXtended LSB (XLSB) methods [18] for all audio files. XLSB was presented and evaluated in [18]. While performing a T-Test between PSNR values of ALSB and SLSB the result in (p=0.00088), which indicates a significant difference between these PSNR values with an advantage to ALSB. Moreover, the BER result confirmed that the proposed ALSB over performed SLSB and XLSB. ALSB algorithm achieved the lowest BER values comparing to other algorithms. T-test between BER values of ALSB and SLSB results in (p = 0.0000735103), which ensures a significant difference between BER values with an advantage to ALSB. Accordingly, the proposed ALSB achieved high performance and outperformed SLSB algorithms.

Table 2 PSNR and BER results

Name of Audio generator	XLSB PSNR	SLSB PSNR	ALSB PSNR	XLSB BER	SLSB BER	ALSB BER
Pop	67.0086	61.1675	69.0515	0.04	0.05	0.0025
Rock	66.758	61.9193	68.0656	0.038	0.041	0.0024
Blues	67.9554	62.4768	68.5194	0.037	0.04	0.0024
Hip-hop	58.8168	62.8794	59.9845	0.035	0.038	0.0022
Dance	65.2817	62.7883	66.4976	0.034	0.037	0.002
Metal	66.8681	62.9386	67.8172	0.031	0.034	0.0019

Figures 1 and 2 illustrate the PSNR and BER results for XLSB, SLSB and ALSB techniques.

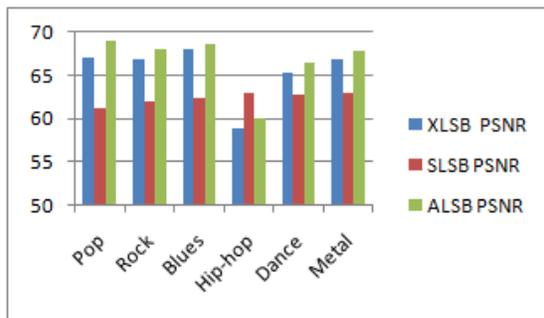


Figure1. PSNR Comparison Results

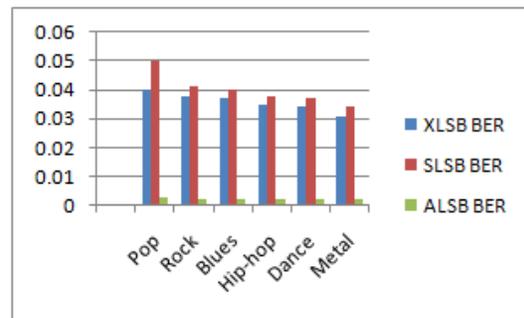


Figure2. BER Comparison Results

5. CONCLUSIONS

This paper has investigated audio steganography, particularly with respect to MP3 files after compression. In this concern, a new algorithm known as Advanced least significant bit algorithm (ALSB) was presented aiming to meet audio steganography requirements including; imperceptibility, capacity, and robustness. The proposed ALSB enhances steganography efficiency, by not embedding the message in every byte of audio file. Alternatively, the location of the first bit to be embedded is selected randomly and remaining bits are embedded considering odd and even byte values in the audio file.

ALSB algorithm is considered an extension of standard least significant bit (SLSB). SLSB holds sufficient information about cover format bits which are not manipulated. This increased errors or distortions. In this work, ALSB was implemented and evaluated with comparison to SLSB. Measurements and results indicated that in ALSB achieved improved capacity and increased PSNR values (as imperceptibility representative) comparing to other methods such as SLSB and XLSB. In addition, ALSB has shown an increased robustness against attacks by applying BER. Accordingly, experiments show that ALSB method achieves increased average of capacity, improved imperceptibility and advanced robustness.

REFERENCES

- [1] Lentij J., "Steganographic Methods", Department Of Control Engineering And Information Technology, Budapest University. Periodica Poltechnica Ser. El. Eng. Vol.44, No. 3–4, P. 249–258 (2000), Url: [Http://Www.Citesseer.Ist.Psu.Edu/514698.Html](http://Www.Citesseer.Ist.Psu.Edu/514698.Html).
- [2] Katzenbeisser S., Peticotas F., "Information Hiding Techniques For Steganography And Digital Watermarking", Artech House Inc.2000.
- [3] Petitcolas F.A, Anderson R.J., Kuhn M.G., "Information Hiding – A Survey", Ieee, Special Issue On Protection Of Multimedia Content: 1062-1078, July, 1999.
- [4] Cacciaguerra S., Ferretti S., "Data Hiding: Steganography And Copyright Marking", Department Of Computer Science, University Of Bologna, Italy, Url: [Http://Www.Cs.Unibo.It/~Scaggiag/Home-File/Teach/Datahiding.Pdf](http://Www.Cs.Unibo.It/~Scaggiag/Home-File/Teach/Datahiding.Pdf).
- [5] Nedeljko C. (2004). Algorithms For Audio Watermarking And Steganography. Acta Universitatis Ouluensis. Series C., 2004..
- [6] Andres G. (2002). Measuring And Evaluating Digital Watermarks In Audiofiles. Washington Dc. 2002.
- [7] Andres G. (2002). Measuring and Evaluating Digital Watermarks in Audio Files. Washington DC. 2002.
- [8] Nematollahi, Mohammad Ali, Chalee Vorakulpipat, and Hamurabi Gamboa Rosales. "Audio Watermarking." Digital Watermarking. Springer Singapore, 2017. 17-38.....
- [9] Shrivastav, Vijay. "A study of different steganographic methods." Journal of Digital Integrated Circuits in Electrical Devices 2.1 (2017): 1-6.
- [10] Arab, Farnaz, and Mazdak Zamani. "VW16E: A Robust Video Watermarking Technique Using Simulated Blocks." Multimedia Forensics and Security. Springer International Publishing, 2017. 193-221.
- [11] Atoum, Mohammed Salem. "A Comparative Study of Combination with Different LSB Techniques in MP3 Steganography." Information Science and Applications. Springer Berlin Heidelberg, 2015. 551-560..
- [12] Atoum, M. S., Suleiman, M., Rababaa, A., Ibrahim, S., & Ahmed, A. (2011). A Steganography Method Based on Hiding secrete data in MPEG / Audio Layer III. Journal of Computer Science, 11(5), 184-188 [12] Deng, K., Tian, Y., Yu, X., Niu, X., Yang, Y., & Technology, S. (2010). Steganalysis of the MP3 Steganographic Algorithm Based on Huffman Coding. Test, (1), 79-82
- [13] L. Maciak And M. Ponniah And R. Sharma, "Mp3 Steganography", 2008.
- [14] Atoum, Mohammed Salem, Subariah Ibrahimn, Ghazali Sulong, Akram Zeki, and Adamu Abubakar. "Exploring the challenges of MP3 Audio steganography." In Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on, pp. 156-161. IEEE, 2013. Bhattacharyya, S., Kundu, A., Chakraborty, K., & Sanyal, G. (2011). Audio Steganography Using Mod 4 Method. Computing, 3(8), 30-38.

- [15] Bhattacharyya, S., Kundu, A., Chakraborty, K., & Sanyal, G. (2011). Audio Steganography Using Mod 4 Method. *Computing*, 3(8), 30-38.
- [16] El-Bendary, Mohsen AM. "FEC merged with double security approach based on encrypted image steganography for different purpose in the presence of noise and different attacks." *Multimedia Tools and Applications* (2017): 1-39.
- [17] AbedulsalamAlarood, Alaa, et al. "HIDING AMessage IN MP3 USING LSB WITH 1, 2, 3 AND 4 BITS." *International Journal of Computer Networks & Communications (IJCNC)* Vol.8, No.3, May 2016.
- [18] Atoum, M.S, " MP3 audio steganography techniqu using extended least significant bit", Thesis (Ph.D (Sains Komputer)) - Universiti Teknologi Malaysia, 2014.

SECURING ONLINE ACCOUNTS VIA NEW HANDSHAKE PROTOCOL AND GRANULAR ACCESS CONTROL

Mehrdad Nourai and Haim Levkowitz

Computer Science Department,
University of Massachusetts Lowell, Lowell, MA, USA

ABSTRACT

When we need to make informed financial decisions, we seek out tools to assist us with managing and aggregating our finances. Traditionally, money management software packages have been available for personal computers; however, they were expensive and often had steep learning curve. With a paradigm shift to cloud-computing users are looking toward the web for an easier and low-cost solution. As a result, third-party companies have been formed to fill this gap. However, users have to share their login credentials with the third-party, and if that information gets compromised, an attacker can access and perform transactions on their account.

We present a novel, holistic model with a new handshake protocol and access control, which authenticates and forms a sandbox around a third-party access. When utilizing these novel techniques, users' original login credentials can remain private, and no one would be able to perform transactions on the users' account.

KEYWORDS

Security, Network Protocols, SSL Cryptography, PKI

1. INTRODUCTION

Today, all of our financial accounts are accessible online, and often they tend to be with different institutions. When one needs to figure out the overall picture of their finances (e.g., net worth or track what is happening with their money), one would need to start by logging into each of their accounts individually. This involves remembering login credentials for each account. Ideally, for good security practices, each account should have unique login credentials, however as a convenience, users' may use one set of credentials for most (if not all) of their accounts. Once the users log into their account, to get a big picture, they would need to download their account information in the proper format and import it to a locally installed financial software package (e.g., Intuit Quicken). Although using these tools are an improvement over tracking your financial life by hand, this process can be tedious, time-consuming, and may become overwhelming for some users that are not familiar with the world of finances. There are usability issues and inconveniences with the locally installed budgeting applications. For instance, the software localized to one computer and needs to be installed, maintained and patched to avoid security vulnerabilities. Also, they tend to be expensive, and their interfaces are a bit complicated to use and often change over time with each iteration or edition of the software. This model has a steep

learning curve and although it may have been sufficient or the only form of financial aggregation software available years ago, it is no longer the case. Thus, it is not surprising that users are migrating to online tools for managing their personal finances.

The idea behind the third-party company is to provide a set of budgeting features that were previously offered by the locally installed financial software, with the added advantage of the third-party software doing all the work for free or for a small fee. For third-party companies to provide these services, they would need to login to users' online accounts to read and collect information. The third-party can utilize desired algorithms on the users' account information to critically examine transactions, net worth, and other finances, then create and present an aggregate report in a textual or graphical manner. This is a preferred method among users, who may have used locally installed software that they had to purchase, keep updated, and perform backups on a regular basis.

Although the online budget aggregate tool is an excellent and affordable tool, users have security concerns and are vulnerable to attack when they sign-up to use these types of services. The vulnerability starts by giving private accounts' login credentials to third-party companies. If an attacker manages to compromise third-party provider's computer, they have got users' entire financial lives in their hands. This is because, the current design of online accounts is in a way that when someone logs into a bank account, everything that the owner of the account can do there, they can do, too. That is, it is not just that they could look at one's bank accounts, or credit card information, they can also transact on it, too. The main idea of this paper is to showcase a novel and holistic login design with techniques for securing online accounts, by leveraging a whole new separate set of login credentials with lower permission. We explain precisely the proposed techniques which are needed to protect online accounts from potential fraudulent activity when users utilize services offered by third-party companies. The main contributions of this paper are:

- To introduce a new handshake protocol to be used by a third-party to authenticate access to users' online accounts (discussed in Section 4.2).
- To introduce a new granular access control layer for fine-grained access capability to users' online accounts (discussed in Section 7.4).

2. EXISTING PRACTICES AND INFRASTRUCTURE SHORTCOMINGS

We now describe what is currently being used in practice and its shortcomings.

2.1. Current practices

For financial institutions to provide a secure online mechanism for their customers to access their accounts, financial institutions utilize HTTPS (HTTP over SSL) technology that can be used via a web browser to access their account. The current process is as follows; one opens a web browser on the user's computer, types in "https://" followed by the website address of their financial institution. This communication between the client and server is secured using encryption and handshaking via SSL protocol. When the page is displayed, it may contain an area within the page to obtain the customer's login credentials or may have a sign-in link to open a login page. The mechanism used for inputting the account credentials utilizes HTML FORM INPUT tag via POST method. Customers get a couple of textboxes, one for the username and one for the password and a sign-in button. Once the user inputs the necessary fields and clicks the sign-in button, the process of user authentication gets started.

The financial institutions' server converts customers' passwords into a custom hash using the algorithms they first used to create the hash (i.e., when the user first set up the account's password), and checks for a match. If it matches, the customer is allowed to access the account. If it does not match, the server may allow a few more tries before locking the account. In addition to this process, financial intuitions often ask for more verification if their customer is signing in from a new device. This extra step involves one or more security questions that the user provided answers to when they first set up their account's credentials. The extra step is an added security measure to ensure that even with the hacked password, the attacker would have one more challenge before gaining access to the account. With the current practices, there are variations to these steps. Hence, not all financial institutions follow a standard mechanism for authenticating their users' access. For instance, some may display a predefined image for identification of the legitimate versus fake website that was made to look like their financial intuition's website. Others may ask for the username on the home page but not password at first, until the user clicks the sign-in, continue, or next button. There is also the second authentication using a code that is delivered either via email, phone, or a mobile app. In general terms, the extra steps may consist of some other information that the owner of the account knows other than their password which financial institutions can ensure it is indeed the owner of the account that is attempting to accessing the account. Therefore, a hacker has to break down at least two barriers to gain access to the account. While this process works well in practice, developers designed it for the humans' capabilities, and not for machines. Thus, financial institutions had to make their interface easy enough for human use, as well as appealing to the masses that use online services. That is, the login credentials used in current practices are a compromise between security and user's convenience.

2.2. Current Infrastructure Shortcomings

The infrastructure of online accounts lacks the mechanisms to support a different form of account authentication with restrictive access. As a result, users are giving their personal access credentials to third-party companies to utilize the financial services they provide. With ever-increasing cyber-security threats, coupled with the existing industry practices, users may make themselves vulnerable to cyber criminals. The current practices have created a challenging and engaging problem that needs to be solved to keep the users safe from potential cyber-attacks.

The following is a list of potential problems with the current infrastructure:

- Users are making themselves vulnerable by giving their banking credentials in the form of username/password plus security questions and answers to third-party companies.
- The users' credentials are designed to be private and not shared with others.
- Once users' credentials are given to a third-party company, they can be stored on their server, which may be exposed to an attacker.
- Users may use the same username/password for other accounts or other places, and as a result, if an attacker steals their credentials, they can access more than just that account.
- Current bank accounts are full-access accounts, and as a result, once a third-party company has access to these accounts, they can perform transactions on that account.
- Financial institutions are not the only company that always allow full-access once users' credentials are entered. Hence, other online accounts that users share with others may be at risk of being vulnerable to an attacker.

3. NETWORKING INFRASTRUCTURE

In this section, we will discuss the foundation of the networking infrastructure that our new protocol will utilize to deliver the needed security.

3.1. Secure Channel

Secure channels between two parties are needed when the transmitted information is sensitive and private while traveling over an insecure medium (e.g., the Internet). The current practices referred to as SSL, which is for securing a channel for private communications will be discussed next.

3.2. Secure Sockets Layer

The current secure connection technology used on the Internet for securing communications between a client and a server is called SSL (Secure Sockets Layer), and its predecessor TLS (Transport Layer Security). Although TLS is the next generation of SSL, the term SSL is prevalent and therefore we will use it throughout this document.

SSL protocol was originally developed by Netscape Communications in 1994 to address security issues of communicating via the Internet [1]. The protocol was a revolutionary technology for securing the Internet traffic that carried personal or sensitive information. The SSL technology is over two decades old and has evolved over time to be more secure. When new SSL vulnerabilities are discovered, computers become faster, and security demand of institutions grows, SSL will continue to evolve over time to address the needs of users. The workings of SSL depend on trust models provided by Certificate Authorities (CA) and public key infrastructure which is based on cryptography. Its underlying mechanisms protects the transmitted information integrity and security by providing authentication of the server to the client, and optionally provides authentication of the client to the server. Although SSL has several security features built-in, it is not immune to cyber-attacks (discussed in Section 8.2). Our new protocol will leverage SSL technology and its ability to secure the communications between client and server without any changes to this layer.

3.3. Transport Layer

The Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are located in this layer. TCP is a connection-oriented protocol and has three-way handshakes to establish the connection between the client (Initiator) and server (Receiver). TCP is the most reliable and prevalent protocol on the Internet because it guarantees packet delivery, ordering, and congestion control. UDP is a connection-less protocol that does not guarantee packet delivery and ordering which is typically used for streaming applications. The TCP along with IP layer, make up the TCP/IP protocol, which we will use as our transport layer protocol. No major changes are needed in this layer, however, with a minor exception that TCP flags might be set to flush out the packets.

3.4. Secure versus Insecure Network Ports

The TCP and UDP transport layer protocols have 65535 ports each. The Internet Assigned Numbers Authority (IANA) assigns the ports to be used for specific protocols and general use [2]. Although some ports are used for secure protocol, there is no such thing as “secure” or “insecure” port numbers. The traffic that flows through network ports can be encrypted or in plain text. Hence it is up to the underlying protocol to use them as “secure” or “insecure” ports.

Nevertheless, there are benefits in standardizing assignments of default ports for “secure” or “insecure” traffic. This may reduce confusion or errors of using certain common ports for secure communications during the setting up of the firewall rules.

4. APPLICATION LAYER PROTOCOLS

In this section, we will discuss the current HTTPS and its issues, and then present our new HTTPAS protocol which addresses security concerns of current practices when used with third-party companies.

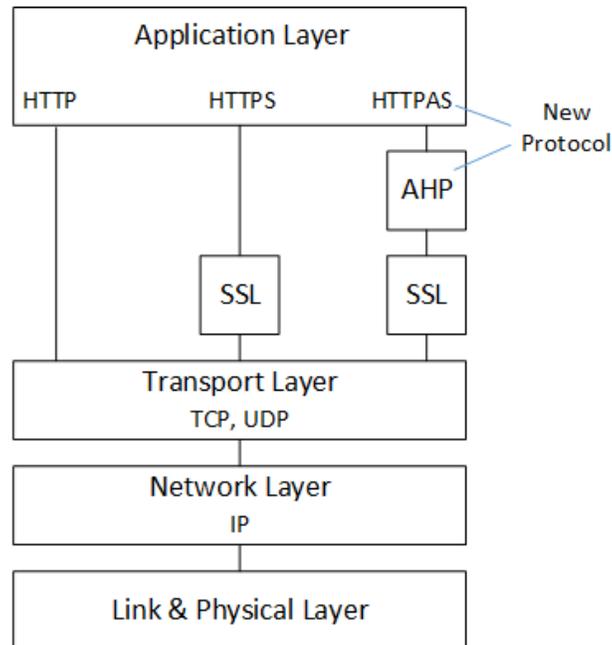


Figure 1. The TCP/IP stack with the new HTTPAS and the new authentication handshake protocol

4.1. HTTPS Protocol

The Internet is an open and insecure medium for communication, hence, when a connection is made between a client machine and a server, all the information transferred over the Internet is traveling over an insecure wire. This information can be intercepted and be seen by others. Therefore, sensitive information (e.g., username/password, bank account information, medical records) must be protected while the information is in transit. The Hypertext Transfer Protocol (HTTP) that is mostly used for accessing information via the web is done in plain text which makes it vulnerable to an attacker. To protect data going between a client and a web server, a protocol called HTTPS is used. HTTPS consists of HTTP over SSL protocol, which encrypts and secures the communication to protect it while in transit [3]. HTTPS works flawlessly and behind the scene, and it does that without user intervention unless something goes wrong with the connection, which then informs the user and allows the user to decide what to do next. When users open a website using HTTPS with their browser, and if it supports the HTTPS protocol, a green lock icon is shown as part of the link in the browser. Once the lock is clicked, information about the security and connection are displayed. For instance, details can show the connection information as TLS 1.2 AES 256 GCM RSA (2048), which is the version and cryptography specifications of the connection. HTTPS was geared toward human usage for securing their sensitive online web activity and interactions. No changes are needed to this protocol. Instead, we

will introduce a novel protocol for computer-to-computer communications which we will discuss next.

4.2. Introducing the New HTTPAS Protocol

We have designed and created an architecture for a novel application layer protocol called HTTPAS, which stands for HTTP with new Authentication Handshake Protocol (AHP) over SSL. This new protocol utilizes SSL to secure a communication channel between a third-party's computer and a financial institution's web server and then uses the new AHP for the two computers negotiate and authenticate secure access to users' accounts. The motivation for a new protocol is flexibility, extra security, and custom enhancements that the current HTTPS does not offer. The HTTPS protocol was designed to be a general multipurpose protocol for providing secure communication channel on the web. This protocol is often used for human-to-computer communications which require more conveniences for a human user. Therefore, security versus convenience became a compromising factor.

The new HTTPAS protocol closes this gap by offering a set of features that is well-suited for computer-to-computer communications. This protocol can also be adapted for human-to-computer communication, however, due to extra security, it would require more efforts on the human side. Our new approach increases the security to another dimension by utilizing a public key infrastructure (PKI) framework. As a result, we can eliminate the need for third-party companies to ask for and use a user's username/password plus other login credentials while offering extra and better security not found in current practices. We will explain components of this new protocol (discussed in Section 5) in greater details later in the paper.

The diagram in Figure 1 shows HTTPAS within the realm of the TCP/IP network stack. The new protocol consists of new Authentication Handshake Protocol (AHP) (discussed in detail in Section 5), which authorizes a client computer and authenticates access to users' accounts. It uses SSL as its foundation for a secure communication channel. Using existing SSL protocol will reduce or minimize the risk of introducing new vulnerabilities.

The following are the benefits of using HTTPAS for third-party access:

- The solution we are envisioning would result in better security practices which address concerns of existing users and potential new users. The existing users get the better security. The new users that were hesitant to sign-up with a third-party due to security issues of such services, can now be sure that the third-party cannot perform any transactions on their accounts. This can potentially increase the number of users' base utilizing third-party services which benefit all parties involved, i.e., users, banks, and third-party companies as a whole.
- Users' don't have to give their banking credentials which can be in the form of username/password plus security questions and answers to a third-party site, i.e., their credentials which were meant to be private, can stay private and not shared. Instead, the third-party will utilize the new handshake protocol and access control for accessing users' accounts.
- Often users have the same username/password for more than one online account. As a result, once an attacker steals their credentials from a third-party's server, an attacker can get access to those accounts in other places, which can become a major security issue for the users.

- The solution is not limited to banking websites, hence, it can be adapted for other sites, such as email or any online accounts, in general, that use usernames/passwords for their authentication. A third-party can access these accounts on a read-only basis.
- If a third-party's server is compromised by an attacker and access information is stolen, the attacker cannot perform transactions on the account. In addition, the bank and/or the owner of the account can easily revoke the access and generate a new access protocol, which can be safer and more convenient than with current practices.

4.3. Port Numbers

To place the separation of traffic between the current and the new protocol, as well as, minimize or eliminate any changes to the existing protocols, HTTPAS uses a different TCP/IP port number than HTTPS. The current application layer TCP/IP protocol port assignments for existing application protocols have been designated by IANA, which are as follows: port 80 for HTTP and port 443 for HTTPS. The new HTTPAS protocol does not currently have a designated port assignment. Hence, in the interim, we will use default port 10443 which according to the IANA online registry search tool, has not officially been assigned to any protocol.

5. NEW AUTHENTICATION HANDSHAKE PROTOCOL

The new Authentication Handshake Protocol (AHP) utilizes public key infrastructure (PKI) framework, TCP/IP as its transport layer and SSL as its transport layer security for its underlying mechanisms. AHP is the main component and workhorse behind the new HTTPAS protocol and is responsible for authorizing and authenticating the access to users' accounts. In this section, we will describe the precise details of the new protocol.

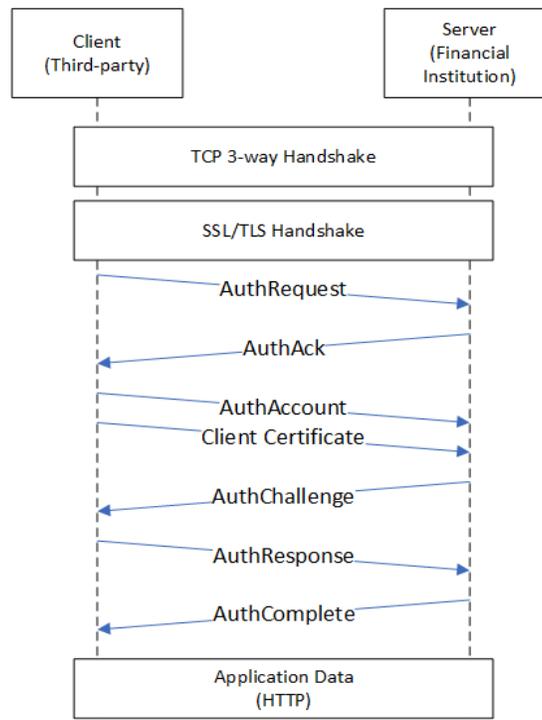


Figure 2. Sequence Diagram

5.1. Cipher Specification

We will use the secure communication channel that SSL provides as the foundation of a secure connection. SSL negotiates Cipher Spec between client and server, therefore, for performance reasons we will leverage the negotiated Cipher Spec to minimize unnecessary handshaking between client and server.

5.2. Sequence of Events

AHP is responsible for the following sequence of Events:

- Negotiate AHP version to use for the handshake
- Obtain user account id that the client is requesting access for
- Obtain client's computer certificate and public key
- Verify client's computer certificate with certificate authorities
- Verify that the user account has been authorized to be access from the client computer
- Authenticate the access utilizing challenge/response and pass-phrase
- Grant or deny access to user's account

5.3. Sequence Diagram

The diagram in Figure 2 shows the new handshake protocol between the client and the server. Note that the TCP and SSL handshake protocols have been drawn as a rectangle on the sequence diagram to show their placement within the protocol time line. A detailed explanation of those protocols is outside the scope of this paper. The new handshake consists of several components as follows:

- AuthRequest
- AuthAck
- AuthAccount
- AuthChallenge
- AuthResponse
- AuthComplete

We now describe each of the new handshake's components in a greater detail.

5.3.1. AuthRequest

The first step of the new AHP handshake protocol is AuthRequest, which occurs right after completion of SSL handshakes protocol securing the channel for communication. The client will

start by sending a request for authentication by providing a list of AHP protocol version numbers it supports in a string (listed in order of preferred version numbers). The string must contain a comma separated displayable characters, with two zeros as terminating character. The purpose of exchanging version numbers between client and server is that in case either of the parties does not support some newer version numbers, the client and server can synchronize on which AHP protocol version they can use for the successful handshake.

5.3.2. AuthAck

The server will check to make sure the version numbers are acceptable and respond to the client with a version number string of its own (similar string format as client version number string). The client adjusts the protocol to the server required version number. Note that, for security reasons, only the server can decide on what AHP protocol version number to use. If the server does not support or agree on the client's desired version number, the connection is terminated.

5.3.3. AuthAccount

The client provides user account id information along with its CA certified digital certificate and public key. Once the server receives the necessary information, it checks the client's certificate with certificate authorities for authenticity and extracts domain information. Now the server must check to ensure that the user account id exists and the client has been authorized to access the account. The connection will be terminated if any of these steps fails to complete successfully. Note that although SSL provides an optional client verification, we will perform the client verification as a mandatory step here. This extra step ensures that even if the user account's key were compromised, the key would not be accepted from an unauthorized client.

5.3.4. AuthChallenge

The server creates a challenge string made up of an arbitrary length of random numbers, and then encrypts it with the user id's public key and sends the challenge string to the client. The random number can prevent a forward attack at which the challenge string cannot be reused for future authentication.

5.3.5. AuthResponse

The client must first decrypt the user account's private key utilizing a pass-phrase, then decrypt the challenge string and send it to the server.

5.3.6. AuthComplete

The server checks the decrypted string, and if there is a match, it returns AHP_SUCCESS. Otherwise, it returns an AHP_FAILED code. If authentication was successful, the HTTP request/response commands could be sent, otherwise the connection will be closed and no other data would be accepted.

6. TRUST MODEL

To identify a server or a client machine that we want to communicate with, a trusted source is needed to verify the identity of either of the party. We will discuss that next.

6.1. Certificate Authority

Certificate Authority (CA) is an entity (e.g., VeriSign, GeoTrust, Comodo, DigiCert) that is trusted and which issues digital certificates to be used for authentication of computer systems on the Internet. CA validates the originator of the request for a digital certificate before issuing their certificate. The concept is similar to trusting government institutions that issue a driver's license which is often used for identification.

6.2. Certificate

Certificates are based on X.509 standard [4] and are digital documents which are generated and issued by a CA. Certificates are made utilizing the public-key cryptography (PKI) which enables parties communicating over the Internet/network to validate the identity for establishing a secure connection [5]. Digital certificates can be purchased from CAs for an annual fee or other purchase plans offered by a particular CA. Certificates can also be locally generated (i.e., self-signed certificate) which are typically used for development and testing purposes. We require that third-party companies purchase digital certificates from their desired CA and deploy them on their computers to be used with our handshake protocol.

6.3. Certificate Revocation Lists

To identify the validity and status of a digital certificate (e.g., revoked or expired), CA keeps a record of those certificates in a list called the Certificate Revocation Lists (CRL) [6]. Our handshake protocol will check the CRL to ensure that the digital certificate is valid and has not been revoked. This is a critical step in the authentication process to ensure that the client or server meets the identity requirements of the CA.

7. ACCESS CONTROL

In general terms, access control includes the systems protection mechanisms for granting/denying access to the available resources. This ensures that persons or services that are requesting access to a resource have been authorized prior to use of the resource. The resource can have (but not limited to) the following protection attributes: read, write/modify, and delete. The protection mechanism needs to be designed for each type of system and may vary widely from system to system. [7]

7.1. Principles of Least Privilege

The best approach to giving authorization to a resource is to use the concept of giving the Least Privilege, i.e., giving what is needed to use a resource and restricting everything else [8]. Utilizing this principle can limit what can be done to a resource, which can minimize data leaks or misuse.

7.2. Access Models

Access control security can be designed for infrastructure protection via one or more of the following security models [9] (listed in alphabetical order):

- Access Matrix Model
- Bell-LaPadula Confidentiality Model

- Clark-Wilson Integrity Model
- Discretionary Access Control (DAC)
- Mandatory Access Control (MAC)
- Role-Based Access Control (RBAC)

These security models provide an array of protection and access control for a variety of real-world applications, such as operating systems, database, and web systems. This paper is interested in the Role-Based Access Control (RBAC) model since it provides the necessary granular access for securing online accounts. The RBAC has been used for protecting a variety of domains for government and commercial infrastructure [10]. We will utilize the RBAC to define new roles and types of access for particular web objects to provide added security for online accounts.

7.3. Current Access Mechanism

With the current account access mechanism, when a third-party company logs into users' online accounts at their financial institutions, they get the same privileges as the account's owner, i.e., full privilege access to users' accounts. This is the by-product of using the same authentication mechanism for logging into a user account, which makes it difficult to distinguish access made by a third-party company versus the account's owner. As a result, financial institutions may not be able to offer any account access control to their users, unless they can differentiate between accesses made to the same account. We believe that no one other than the account's owner should have full privileges to users' accounts, even when they were logged in successfully to that particular account. Therefore, the third-party access must be limited and sandboxed to only allow display of account information or authentication of the account, without any other privileges which account owner usually gets.

7.4. Granular Access Control via RBAC Model

With the new online account security model, we are envisioning in this paper, it makes it feasible to offer granular access control for the online accounts. Since it would be possible to distinguish third-party accesses from the account's owner access (e.g., based on which protocol used). Utilizing a granular access control with the role-based scheme of the RBAC model enables fine-grained access to sandbox third-party companies. When a third-party uses the new protocol to access the users' accounts, the source login by definition becomes distinguishable, and sandbox becomes feasible. That is, financial institutions can detect the alternative form of account access, when it is initiated using the new protocol versus current protocol, and then allow the access according to the initiator's role.

We now define the architecture of the access control structure as access pair (Object, Attribute) with Role in greater detail.

7.4.1. Roles

We define the following three roles to address the security of the accounts and enable each role to perform their authorized access:

- **ROLE_ADMIN** - This role performs account administration tasks. It offers full privileges that can allow an account manager at a financial institution to make changes to users' accounts.

- **ROLE_OWNER** - This role gives privileges to the owner of the account, which includes reading the contents of the account and performing transactions on the account.
- **ROLE_THIRDPARTY** - This role is used by third-party companies that are authorized by users to read the contents of their accounts. No transactions are allowed on the account via this role.

7.4.2. Object

The Object is the entity that needs to have granular access protection which can apply to the whole page, individual accounts, any components within of the page, to name a few. For example, objects can be a Checking Account, a Savings Account, or a Certificate of Deposit.

7.4.3. Attribute

The Attribute is a word length which consists of fine-grained privileges allowed for a particular object. We define four Attribute flags in a hexadecimal nibble format and present it as binary numbers as shown in Table 1:

Table 1. List of Attribute flags and their values.

Attribute flag	Value (binary)
ATTRB_READ	1000
ATTRB_MODIFY	0100
ATTRB_TRANSACT	0010
ATTRB_CUSTOM	0001

The binary numbers can be used with the logical OR operator to build a hexadecimal nibble representing an access configuration for a particular role. Then each role will have its dedicated hexadecimal nibble for specifying the privileges for that role. Note that, for the majority of objects, the first three privileges should cover most of the permission cases. However, a “custom” attribute has been added in case there are particular circumstances that an extra permission is needed.

7.4.4. Role Mask

The least significant hexadecimal nibble of an Attribute word is assigned to **ROLE_ADMIN**, then moving toward the most significant bit, the next nibble is assigned to **ROLE_OWNER**, and the next is assigned to **ROLE_THIRDPARTY**. We define role mask for the Attribute word in hexadecimal number as shown in Table 2:

Table 2. List of Role masks and their values.

Role mask	Value (hexadecimal)
ROLE_ADMIN_MASK	F
ROLE_OWNER_MASK	F0
ROLE_THIRDPARTY_MASK	F00

If new roles are needed, it can be directly added to the Attribute word after the last role moving toward the most significant digit. For example, access pair (Checking Account, 0x8EC) means, third-party can read, but cannot perform transactions or modify the Checking account in any shape or form; owner of the account can read, modify, and transact on the account; administrator

(e.g., manager) can read and modify, but cannot perform any transactions on the Checking account.

8. SHORTCOMINGS

In this section, we will discuss security and performance shortcomings of our new techniques that we have discussed in this paper.

8.1. Performance Shortcomings

The new protocol HTTPAS is more secure when used for computer-to-computer communications. However, this new security enhancement comes at the cost of performance. The shortcomings of performance are mainly due to two major security steps such as verifying client's certificate and using public key infrastructure which uses asymmetric cryptography. Asymmetric encryption/decryption utilizing larger bits to avoid brute-force or other forms of attacks will run slower than a simpler method with lower security cryptography.

8.2. Security Shortcomings

In today's modern connected world, with many types of devices connecting us all together, security has never been more important. During our research and writing of this paper, cyber-attacks on two major companies were made public. Yahoo email attack was made public with data stolen from one billion accounts [11] and Yahoo data being sold on the "dark web" [12]. Dyn (domain name service provider) was attacked via Distributed-Denial-Of-Service (DDOS) causing interruption of services to the popular websites [13].

We now discuss security issues and vulnerability that can affect the security and integrity of secure communications over the Internet/network in greater detail.

8.2.1. SSL Protocol Vulnerability

Although SSL design and architecture has built-in mechanisms for preventing attacks (e.g., Eavesdropping and Man-In-The-Middle attack prevention features), it is not immune to attackers [14]. The vulnerability may come from flaws in SSL protocol, SSL libraries, hacked digital certificates, and cipher negotiations to name a few. These flaws are often discussed at security conferences such as Black Hat. Our protocol will be vulnerable to attackers if these infrastructures that we utilize as a foundation for our protocol become vulnerable.

8.2.2. CA Entities Vulnerability

The CAs are the foundations for trust model of digital certifications. If a vulnerability exists in CAs infrastructures and attackers can exploit it to their advantage, it will break the trust models. This may have such an adverse effect on the reputation of the CA which may result in cease of the operation of the CA. Nevertheless, if CA or certificates are hacked, it will make our protocol vulnerable to an attacker. An example of an attack on CA is, a CA named DigiNotar which was hacked by an attacker in 2011. The attack on DigiNotar had a catastrophic effect, and as a result, they were not able to recover from the attack, and filed for bankruptcy [15], [16].

8.2.3. Cryptography Vulnerability

Encryption algorithms have been proven to be mathematically sound, however, as computer systems are getting more powerful over time, the number of bits used in calculations of the

encryption mechanism must also increase. For example, RFC7935 states that the RSA key pairs for use in public key infrastructure (which is a framework for what is called one-way function) must be using at least 2048-bit [17]. Therefore, the strength of encryption algorithms for a one-way function is based on the computing power and time required to decrypt the message.

8.3. Initial Burden for all Parties Involved

All parties involved (e.g., financial institutions, third-party companies, end-users) would need to make changes to their current processes, practices, and habits. For instance, financial institutions would need to modify their online accounts and systems to use the new protocol and offer new account granular access control; third-party companies must make changes to their current informational retrieval processes, as well as follow the specification of the new protocol for user authentication; the end-users would need to request alternate access credentials from their financial institutions for their accounts, provide the vendor information of the third-party, as well as change their full-access login credentials if they have already exposed that to others.

9. RELATED WORK

In this section, we will discuss related work in the area of authenticating users' access to online services.

9.1. OAuth

A related work to our research is the OAuth authorization framework. OAuth allows authentication of a user without exposing their account credentials to a third-party provider over HTTP protocol [18]. The main concept behind OAuth is that, when a user is trying to login to a third-party website, they can use their accounts' username/password from one of the major companies (e.g., Facebook, Twitter, LinkedIn) to authenticate themselves with a third-party company. With this method, users no longer have to reveal their login credentials with third-party companies. Currently, the deployments of OAuth framework have had limited exposure. This model has not been widely accepted as the alternative to hiding and protecting account credentials from third-party companies. As a result, OAuth currently suffers from penetration and adoption challenges, as well as privacy and security concerns.

9.2. Application Programming Interface

Another related work is where companies provide mechanisms for information retrieval and manipulation from their systems. This type of access is performed via application permission model [19] provided by Application Programming Interface (API) (e.g., Twitter API) [20]. This method is used to read, write, perform information retrieval and data mining to access and harvest data from companies (e.g., Twitter, Google) that support those APIs. The API may use OAuth technology as its underlying authorization mechanism to authenticate the request. Due to the nature and exposure of information provided with this model, it has its limit where sensitive information, security, and privacy are concerned.

10. CONCLUSIONS

The ubiquity of the Internet has increased the potential exploitation of security weaknesses and vulnerabilities of our financial institutions. Users are sharing their full access accounts' credentials with third-party companies and need to be wary of sharing this sensitive information with others, especially when those credentials can be used by others to execute transactions on

their account. However, users' growing need of financial aggregation services from third-party providers, coupled with the lack of an alternate mechanism for account authentication and lack of restricted access control, can make users' vulnerable to attackers if their access credentials get compromised.

In this research paper, we have introduced and prescribed a novel and holistic model with a new protocol for online account architecture to be used by third-party companies. This model works on the notion of two security components: 1) Authentication mechanism utilizing new handshake protocol which enables verification of users' credentials that is more secure than a username/password combination; 2) User access sandboxing technique via granular access control that protects the account against unwanted transactions. Utilizing new architecture, design, and novel techniques we have presented in this paper, users no longer need to give out their full access accounts' credentials to third-parties. Instead, users can give limited alternate access login credentials for third-party use, and if attackers compromise their computer and access information is stolen, attackers cannot perform transactions on the account. In the case of a security breach, the alternate login credentials can be revoked and reissued with no impact on the existing full access account credentials. Furthermore, our novel and holistic techniques are universal and can be adapted for other domains (e.g., medical records, airline ticket system, online stores, emails) with little or no modifications to our architecture we have presented in this paper.

REFERENCES

- [1] I. Jinwoo Hwang, "The Secure Sockets Layer and Transport Layer Security," Jun. 2012. [Online]. Available: <http://www.ibm.com/developerworks/library/ws-ssl-security/index.html>
- [2] "Service Name and Transport Protocol Port Number Registry," 00017. [Online]. Available: <http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml>
- [3] E. Rescorla, "HTTP Over TLS," Internet Requests for Comments, RFC Editor, RFC 2818, May 2000. [Online]. Available: <http://www.rfc-editor.org/rfc/rfc2818.txt>
- [4] "ITU-T The Directory: Public-key and attribute certificate frameworks." [Online]. Available: <http://www.itu.int/itu-t/recommendations/rec.aspx?rec=13031>
- [5] S. Chokhani, W. Ford, R. Sabett, C. Merrill, and S. Wu, "Internet X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework," Internet Requests for Comments, RFC Editor, RFC 3647, November 2003.
- [6] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk, "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) profile," Internet Requests for Comments, RFC Editor, RFC 5280, May 2008, <http://www.rfc-editor.org/rfc/rfc5280.txt>. [Online]. Available: <http://www.rfc-editor.org/rfc/rfc5280.txt>
- [7] A. Silberschatz, P. B. Galvin, and G. Gagne, Operating system concepts, 9th. Addison-Wesley Reading, 2013.
- [8] G. S. Graham and P. J. Denning, "Protection: Principles and practice," in Proceedings of the May 16-18, 1972, Spring Joint Computer Conference, ser. AFIPS '72 (Spring). New York, NY, USA: ACM, 1972, pp. 417-429. [Online]. Available: <http://doi.acm.org/10.1145/1478873.1478928>
- [9] M. G. Solomon and M. Chapple, Information security illuminated. Jones & Bartlett Publishers, 2009.
- [10] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed NIST Standard for Role-based Access Control," ACM Trans. Inf. Syst. Secur., vol. 4, no. 3, pp. 224-274, Aug. 2001. [Online]. Available: <http://doi.acm.org/10.1145/501978.501980>

- [11] S. Fiegerman, “Yahoo says data stolen from 1 billion accounts,” Dec. 2016. [Online]. Available: <http://money.cnn.com/2016/12/14/technology/yahoo-breach-billion-users/index.html>
- [12] S. Larson, “Hackers are selling Yahoo data on the dark web,” Dec. 2016. [Online]. Available: <http://money.cnn.com/2016/12/16/technology/yahoo-for-sale-data-dark-web/index.html>
- [13] “Hacked home devices caused massive Internet outage.” [Online]. Available: <http://www.usatoday.com/story/tech/2016/10/21/cyber-attack-takes-down-east-coast-netflix-spotify-twitter/92507806/>
- [14] “Logjam: the latest TLS vulnerability explained,” May 2015. [Online]. Available: <http://blog.cloudflare.com/logjam-the-latest-tls-vulnerability-explained/>
- [15] J. Prins and B. U. Cybercrime, “DigiNotar certificate authority breach Operation Black Tulip,” 2011.
- [16] K. Zetter, “DigiNotar files for bankruptcy in wake of devastating hack,” Wired magazine, September 2011.
- [17] G. Huston and G. Michaelson, “The profile for algorithms and key sizes for use in the resource public key infrastructure,” Internet Requests for Comments, RFC Editor, RFC 7935, August 2016.
- [18] “OAuth Community Site.” [Online]. Available: <https://oauth.net/>
- [19] “Application Permission Model - Twitter Developers.” [Online]. Available: <https://dev.twitter.com/oauth/overview/application-permission-model>
- [20] “Twitter Developer Documentation - Twitter Developers.” [Online]. Available: <https://dev.twitter.com/docs>

AUTHORS

Mehrdad M. Nourai received a B.S. degree in Electrical Engineering from Northeastern University, Boston Massachusetts in 1982, and an M.S. degree in Computer Science from Boston University, Boston Massachusetts in 2000. He is currently a Ph.D. degree candidate in Computer Science at the University of Massachusetts Lowell, Lowell Massachusetts. He has over three decades of professional experience in computer industry and academics. His industry experience consists of architect, design, and development of software for all kinds of computer systems including embedded systems and systems with standard, proprietary, and real-time operating systems. He has been teaching computer science courses as an adjunct faculty for the MET Computer Science Department at Boston University since 2000. In addition, he has been teaching courses as an adjunct faculty for the Computer Science Department at Salem State University since 2008. His research and areas of interests include Computer Networks and Security, Data Communications, Human-Computer-Interaction, Database, and Mobile Apps Development.



Haim Levkowitz is the Chair of the Computer Science Department at the University of Massachusetts Lowell, in Lowell, MA, USA, where he has been a Faculty member since 1989. He was a twice-recipient of a US Fulbright Scholar Award to Brazil (August – December 2012 and August 2004 – January 2005). He was a Visiting Professor at ICMC — Instituto de Ciencias Matematicas e de Computacao (The Institute of Mathematics and Computer Sciences)—at the University of Sao Paul, Sao Carlos – SP, Brazil (August 2004 - August 2005; August 2012 to August 2013). He co-founded and was Co-Director of the Institute for Visualization and Perception Research (through 2012), and is now Director of the Human-Information Interaction Research Group. He is a world-renowned authority on visualization, perception, color, and their application in data mining and information retrieval. He is the author of “Color Theory and Modeling for Computer Graphics,



Visualization, and Multimedia Applications” (Springer 1997) and co-editor of “Perceptual Issues in Visualization” (Springer 1995), as well as many papers on these subjects. He is also co-author/co-editor of "Writing Scientific Papers in English Successfully: Your Complete Roadmap," (E. Schuster, H. Levkowitz, and O.N. Oliveira Jr., eds., Paperback: ISBN: 978-8588533974; Kindle: ISBN: 8588533979, available now on Amazon.com:

<http://www.amazon.com/Writing-Scientific-Papers-English-Successfully/dp/8588533979>).

He has more than 44 years of experience teaching and lecturing, and has taught many tutorials and short courses, in addition to regular academic courses. In addition to his academic career, Professor Levkowitz has had an active entrepreneurial career as Founder or Co-Founder, Chief Technology Officer, Scientific and Strategic Advisor, Director, and venture investor at a number of high-tech startups.

INTENTIONAL BLANK

NEW NON-COPRIME CONJUGATE-PAIR BINARY TO RNS MULTI-MODULI FOR RESIDUE NUMBER SYSTEM

Mansour Bader¹, Andraws Swidan², Mazin Al-hadidi¹

¹Department of Computer Engineering,
Al-Balqa'a Applied University, Salt, Jordan.

²Department of Computer Engineering,
Jordan University, Amman, Jordan.

ABSTRACT

In this paper a new Binary-to-RNS converters for multi-moduli RNS based on conjugate-pair as of the set $\{ 2^{n1} - 2, 2^{n1} + 2, 2^{n2} - 2, 2^{n2} + 2, \dots, 2^{nN} - 2, 2^{nN} + 2 \}$ are presented. $2^n - 2$ and $2^n + 2$ moduli are called conjugates of each other. Benefits of Multi-moduli RNS processors are; relying on the sets with pairs of conjugate moduli : 1) Large dynamic ranges. 2) Fast and balanced RNS arithmetic. 3) Simple and efficient RNS processing hardware. 4) Efficient weighted-to-RNS and RNS-to-Weighted converters. [1] The dynamic range (M) achieved by the set above is defined by the least common multiple (LCM) of the moduli. This new non-coprime conjugate-pair is unique and the only one of its shape as to be shown.

KEYWORDS

Binary , Conjugate-Pair , Dynamic range, LCM, Multi-moduli.

1. INTRODUCTION

RNS is known to support parallel, carry-free, high-speed arithmetic , because it is considered as an integer system, that is appropriate for implementing fast digital signal processors [1] . It is also has main importance in Encryption and Cryptography fields. Other applications include – but not limited to - Digital Signal Processing, correlation, error detection and correction [1 - 3].

RNS basis form is a set of relatively prime integers $P = \{ m_1, \dots, m_k \}$ where $\gcd(m_i, m_j) = 1$ for $i \neq j$. In this paper we are showing that the new non-coprime moduli set presented in [2] could be used in the new non-coprime multi-moduli conjugate-pair Weighted-to-RNS converters.

The set P for prime case is the moduli set with the dynamic range (M) of the system $M = \pi m_i$. But for our case and since each conjugate has the number 2 as a common factor other than the number 1 as in the prime one, the $M = \prod_1^k m_i / 2^{(k-1)}$.

For both cases coprime and non-coprime; any integer $x \in [0, M - 1]$ has an RNS representation $X = (x_1, \dots, x_k)$, where $x_i = X \bmod m_i$.

The new thing we come up with here is working with a full non-prime moduli set (i.e. for this case) $\gcd(m_i, m_j) \neq 1$ for $i \neq j$ (1)

RNS systems based on non coprime moduli have also been studied in literature [2] –[5].

Although as discussed in [2] that non-coprime has little studies upon, we still have strong sense that it deserves to work on.

The rest of this paper is organized as follows. In Section 2, overview of the new Non-coprime multi moduli is proposed. Section 3 presents the realization of the proposed forward converter of the new non-coprime conjugate-pair multi-moduli , while the paper is concluded in Section 4.

2. OVERVIEW OF NEW NON-COPRIME MULTI –MODULI

Since almost all previous work stated that [1][3][5] " The basis for an RNS is a set of relatively prime integers; that is :

$$S = \{ q_1, q_2, \dots, q_L \}, \text{ where } (q_i, q_j) = 1 \text{ for } i \neq j \quad (2)$$

with (q_i, q_j) indicating the greatest common divisor of q_i and q_j .

The set S is the moduli set while the dynamic range of the system (i.e. M) is the product Q of the moduli q_i in the set S. Any integer X belonging to $ZQ = \{ 0, 1, 2, \dots, Q-1 \}$ has an RNS representation" .

$$X \xrightarrow{\text{RNS}} (X_1, X_2, \dots, X_L) \quad (3)$$

$$X_i = \langle X \rangle_{q_i}, \quad i = 1, 2, \dots, L \quad (4)$$

Where $\langle X \rangle_q$ is $X \bmod q$.

For our case of **non-coprime** , equation number 4 becomes :

$$X_i = \langle X \rangle_{q_i}, \quad i = 2, 3, \dots, L \quad (5)$$

For both cases (i.e. Coprime and Non-coprime), if X, Y have RNS representations $\{ X_1, \dots, X_M \}$, $\{ Y_1, \dots, Y_M \}$, the RNS representation of $W = X * Y$ (* denotes addition, subtraction or multiplication) is

$$W \xrightarrow{\text{RNS}} \{ W_1, \dots, W_M \}; W_i = \langle X_i * Y_i \rangle_{q_i}, i = 1, \dots, L \quad (6)$$

Another thing to notice here is that our new proposed non-coprime conjugate-pair multi-moduli set is also conjugate even by dividing it by the common factor among its moduli (i.e. number 2 in this case), the shape is to be discussed in another paper. However it has the following form :

$$\{ 2^{n1-1} - 1, 2^{n1-1} + 1, 2^{n2-1} - 1, 2^{n2-1} + 1, \dots, 2^{nN-1} - 1, 2^{nN-1} + 1 \}.$$

The proposed Non-coprime multi-moduli set form is :

$$S = \{ 2^{n1} - 2, 2^{n1} + 2, 2^{n2} - 2, 2^{n2} + 2, \dots, 2^{nN} - 2, 2^{nN} + 2 \}.$$

It is clear that each conjugate-pair on the numbers line is 4 spaces apart. As discussed in [2] it was shown that having the shape of the new non-coprime moduli set (i.e. $\{ 2^n - 2, 2^n, 2^n + 2 \}$) being 4 spaces apart from each other helped in the Forward conversion process (FC) of the moduli. The same space for our new non-coprime multi-moduli is useful indeed.

Lets take an example to show what is meant by the spaces above.

Ex.1 Let $n1 = 3$, $n2 = 4$ for the set S.

Then the set $S = \{ 6, 10, 14, 18 \}$.

Numbers (6 , 10) and (14 , 18) are 4 spaces from each other on the numbers line. This is true for any value taken for $\{ n1, n2 \dots , nN \}$, notice that $n1 < n2 < \dots < nN$; $n1 \geq 2$.

This is need for the management process to prepare the multi-moduli in good shape.

Least Common Multiple (LCM) is must be used for the non-coprime case, since there is a common factor among the modulus numbers.

3. NEW NON-COPRIME MULTI-MODULI PROPOSED FORWARD CONVERTER

This section is preferred to be divided into two sections in order to show simply how it works. Then from the multi-moduli shape provided it is generalized for size of (N).

In the first section, we are going to take $N = 2$, thus the multi-moduli would consist of 4 modulus values. The second sub-section we are having $N = 3$, so there would be 6 modulus inside the multi-moduli set. Forward conversion (i.e. Binary to RNS) is to be implemented for each case.

3.1 VALUES SET OF 4-MODULUS

The multi-moduli set will be of the form, when we take $N = 2$:

$$S = \{ 2^{n1} - 2, 2^{n1} + 2, 2^{n2} - 2, 2^{n2} + 2 \}.$$

If we take as the first example showed $n1 = 3$, $n2 = 4$. The shape of the set was :

$$S1 = \{ 6, 10, 14, 18 \}.$$

M (i.e. Dynamic Range of the set) is calculated through the LCM. For the set S1 it is equal to

$$6 * 10 * 14 * 18 / 2^{L-1}, \text{ where } L = \text{the size of the set.} \quad (7)$$

For this case $L = 4$, so $M = 1890$.

That means any number in the range $[0 - 1889]$ has a unique representation among the proposed set. This dynamic range is larger than the range for $\{2^{n1} - 2, 2^{n1}, 2^{n1} + 2\}$ which equals 120. i.e. $1890 \gg 120$.

It is also having a larger range than the set $\{2^{n2} - 2, 2^{n2}, 2^{n2} + 2\}$ which has $M = 1008$. i.e. $1890 > 1008$.

This is due we are working with 4-moduli set rather than 3-moduli set, and by neglecting the middle modulus (i.e. 2^n) and having the conjugate of $n1, n2$ instead. Mathematically it could be shown as :

$$M1 = 6 * 8 * 10 / 4, M2 = 14 * 16 * 18 / 4 \text{ while } M3 = 6 * 10 * 14 * 18 / 8.$$

Take for $M2$ case, as it has larger numbers than $M1$, $16 / 4 < 6 * 10 / 8$ or by having $6 * 10 / 2 = 30$, $30 > 16$ when comparing them divided 4 (i.e. having a common base of comparison).

The conversion process works as the follow, each modulus having the shape $2^n - 2$ goes to Converter number 1, while the $2^n + 2$ goes to Converter number 2 that works in parallel.

Converter 1 does its work just as figure 1 in [2] showed, figure 2 in the same paper shows converter 2 work.

Hardware implementation for each case is shown in figures 3, 5 of [2].

3.2 VALUE SET OF 6-MODULUS

When we take $N = 3$, then the multi-moduli set will be on the form :

$$S = \{ 2^{n1} - 2, 2^{n1} + 2, 2^{n2} - 2, 2^{n2} + 2, 2^{n3} - 2, 2^{n3} + 2 \}.$$

If we take $n1 = 3, n2 = 4$ and $n3 = 5$ for simplicity. The shape of the set is :

$$S2 = \{ 6, 10, 14, 18, 30, 34 \}.$$

M (i.e. Dynamic Range of the set) is calculated through the LCM. For the set $S2$ it is equal to

$$6 * 10 * 14 * 18 * 30 * 34 / 2^{L-1}, \text{ where } L = \text{the size of the set.} \quad (8)$$

For this case $L = 6$, so $M = 481950$.

That means any number in the range [0 – 481949] has a unique representation among the proposed set. This dynamic range is larger than the range for $\{ 2^{n1} - 2, 2^{n1}, 2^{n1} + 2 \}$ which equals 120.

i.e. $481950 \gg 120$.

It is also having a very large range than the set $\{ 2^{n2} - 2, 2^{n2}, 2^{n2} + 2 \}$ which has $M = 1008$.

i.e. $481950 \gg 1008$.

Finally it has larger range than the set $\{ 2^{n3} - 2, 2^{n3}, 2^{n3} + 2 \}$ which has $M = 8160$.

i.e. $481950 \gg 8160$.

This is due we are working with 6-moduli set rather than 3-moduli set for each case, and by neglecting the middle modulus (i.e. 2^n) and having the conjugate of $n1$, $n2$ and $n3$ instead.

Mathematically it could be shown as :

$M1 = 6 * 8 * 10 / 4$, $M2 = 14 * 16 * 18 / 4$, $M3 = 30 * 32 * 34 / 4$ while

$M4 = 6 * 10 * 14 * 18 * 30 * 34 / 32$.

Take for $M3$ case, as it has the largest numbers than $M1$ and $M2$, $32 / 4 < 6 * 10 * 14 * 18 / 32$ **or** by having $6 * 10 * 14 * 18 / 8 = 1890$, $1890 \gg 16$ when comparing them divided 4 (i.e. having a common base of comparison).

The conversion process works as the follow, each modulus having the shape $2^n - 2$ goes to Converter number 1, while the $2^n + 2$ goes to Converter number 2 that both works in parallel.

Converter 1 does its work just as figure 1 in [2] showed, figure 2 in the same paper shows converter 2 work.

Hardware implementation for each case is shown in figures 3, 5 of [2].

4. CONCLUSIONS

A new non-coprime multi-moduli set has been proposed. A general formula for the dynamic range of it was derived. Algorithm of the special non-coprime multi-moduli set has been suggested. Also a new mathematical algorithm for the new non-coprime multi-set has been proposed.

This research revealed that non-coprime moduli set may be suitable for wide variety of cases not limited to co-prime only (i.e. Conjugate in multi-moduli).

ACKNOWLEDGEMENTS

The author would like to thank everyone.

REFERENCES

- [1] A. Skavantzios ; Y. Wang, "New efficient RNS-to-weighted decoders for conjugate-pair-moduli residue number systems", IEEE Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers, 1999..
- [2] Mansour Bader, Andraws Swidan, Mazin Al-Hadidi and Baha Rababah, "A binary to residue conversion using new proposed non-coprime moduli set", Signal & Image Processing : An International Journal (SIPIJ) Vol.7, No.3, June 2016 .
- [3] A. Skavantzios ; Yuke Wang, "Application of new Chinese Remainder Theorems to RNS with two pairs of conjugate moduli", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM 1999). Conference Proceedings, 1999.
- [4] M. Abdallah, A. Skavantzios, "On the binary quadratic residue system with non coprime moduli", IEEE Trans. On Signal Processing, vol. 45, no. 8, pp. 2085-2091, Aug. 1997.
- [5] Y. Wang, "New Chinese Remainder Theorems", Proceedings of the Thirty Second Asilomar Conference on Signals Systems and Computers, pp. 165-171, 1998-Nov.
- [6] A. Skavantzios, M. Abdallah, "Implementation Issues of the Two-Level Residue Number System with Pairs of Conjugate Moduli", IEEE Trans. On Signal Processing, vol. 47, no. 3, pp. 826-838, March 1999.
- [7] A. Skavantzios, M. Abdallah, "Novel Residue Arithmetic Processors for High Speed Digital Signal Processing", Proceedings of the Thirty Second Asilomar Conference on Signals Systems and Computers, pp. 187-193, 1998-Nov.
- [8] A. Skavantzios ; T. Stouraitis, "Grouped-moduli residue number systems for fast signal processing", IEEE International Symposium on Circuits and Systems, 1999. ISCAS '99.
- [9] R. Katti, "A new residue arithmetic error correction scheme", IEEE Transactions on Computers, vol. 45, no. 1, January 1996.
- [10] Y. Wang, M. N. Swamy, O. Ahmad, "Three number moduli sets based residue number systems", 1998 IEEE International Symposium on Circuits and Systems, 1998.

AUTHORS

Mansour Bader holds a MSc in computer engineering and networks, University of Jordan, Jordan, 2016. BSc Computer Engineering, Al-Balqa Applied University, Jordan, 2008. He is a technical support engineer of computer networks at computer center of Al-Balqa Applied University for 8 years and a half.



Dr. Andraws I. Swidan was born in Al-Karak Jordan in 1954. He received his diploma in Computer Engineering (with honours) and Ph.D. in Computer Engineering from LETI Ulianov Lenin, Sanct Peterburg (Leningrad), Russia in 1979 and 1982 respectively. He Joined the Electrical Engineering Department at the University of Jordan in 1983 and was one of the founders of the Computer Engineering Department at the University of Jordan in 1999. Since then he is a professor at the department. He is also an Adjunct Professor with the ECE department of the McGill University, Montreal, Canada. He holds several technical certifications among which the CISSP. He is an IEEE member, Jordanian Engineering Association member Quebec College of engineers member. He is a Canada Professional Engineer (The province of Quebec). He was member of several national and international scientific committees. He holds several patents and tens of publications. His main areas of research interest are: computer arithmetic, computer security, encryption algorithms.



Mazin Al-hadidi has PhD. in Engineering Science (Computing Machines, Systems and Networks), Institute of Problems of Simulation in Power Engineering Academy of Science, Ukraine/Kiev .1990-1994, with grade Excellent. Bachelor and Master Degrees in Engineering (Computer and intellectual systems and networks) Kiev Institute of Civil Aviation Engineers, as a governmental scholarship, Soviet Union / Kiev, 1984-1990, with grade very good. General Secondary 12 Years Certificate in the Science branch, Jordan/Al-Salt, 1984, with grade very good.



INTENTIONAL BLANK

TOWARDS A MULTI-FEATURE ENABLED APPROACH FOR OPTIMIZED EXPERT SEEKING

Mariam Abdullah¹, Hassan Nouredine¹, Jawad Makki¹, Hussein Charara¹, Hussein Hazimeh², Omar Abou Khaled², Elena Mugellini²

¹Lebanese University, Beirut, Lebanon

²HES-SO/FR, Fribourg, Switzerland

ABSTRACT

With the enormous growth of data, retrieving information from the Web became more desirable and even more challenging because of the Big Data issues (e.g. noise, corruption, bad quality...etc.). Expert seeking, defined as returning a ranked list of expert researchers given a topic, has been a real concern in the last 15 years. This kind of task comes in handy when building scientific committees, requiring to identify the scholars' experience to assign them the most suitable roles in addition to other factors as well. Due to the fact the Web is drowning with plenty of data, this opens up the opportunity to collect different kinds of expertise evidence. In this paper, we propose an expert seeking approach with specifying the most desirable features (i.e. criteria on which researcher's evaluation is done) along with their estimation techniques. We utilized some machine learning techniques in our system and we aim at verifying the effectiveness of incorporating influential features that go beyond publications.

KEYWORDS

Entity retrieval, Expert seeking, Academia, Information extraction

1. INTRODUCTION

Currently, the Web is in the state of always being active since lots of data are being uploaded constantly. As a result, this has caused the Web to witness so much interactions between different organizations (company, university...etc.) for they weren't satisfied with their own knowledge. For this reason, they used the Web to connect to the outside world for development and improvement (socially, scientifically...etc.). However, despite the interactions' usefulness, they couldn't help but drown the Web with plenty of data, recognized as Big Data which is a very common term nowadays. Retrieving information from the Web is classified as non-trivial for this data is likely to contain noise with no guarantees of good quality. One of the things that's been frequently searched for is experts; seeking for experts is defined as the task of taking the user's query as input and generating a ranked list of expert researchers as output. The query denotes the topic of expertise and the generated list is sorted according to their expertise levels in what concerns the query topic. Fifteen years ago, the scientific community showed its interest in this

task and since then, they became highly dedicated to this domain. In spite of the significance this task possesses, some still wonder “wouldn’t it be easier if we counted on human recommendations?” Yet, the fact human judgments might not be based on reasonable criteria answers the aforementioned question.

To estimate a scholar’s expertise degree, the key idea is the set of characteristics on which the evaluation is done, i.e. features and the way of estimating them where recent works have focused on scholar’s academic publications to extract different features in addition to detecting his relations. As a matter of fact, none took notice of the activities a scholar has done beyond publishing, for instance being in a conference committee, being honored and awarded, his seminars and courses, etc... Incorporating these supplementary features means dealing with more data which sounds a bit more challenging, because of the formerly stated data issues. We aim at verifying that going beyond publications enhances the expertise retrieval performance. Among the non-traditional features, we picked the conference committee evidence, because we believe it is substantial to study how good a scholar is at assessing other scholars’ academic works. Concerning a scholar’s publications, we will be considering the conferences’ ranks in which they were published for experts usually publish their valuable work in top-ranked conferences. The main contributions of this work can be summarized as follows.

- We developed an expert seeking system by combining traditional and new features with the use of supervised and unsupervised learning.
- Incorporation of conference committee memberships of scholars. To the best of our knowledge, we are the first to deal with beyond publication features.

This paper is organized as follows. In section 2, we review the related works including expertise evidences, expertise retrieval models, approaches that utilized these models, then we give a brief discussion. In section 3, we present the proposed approach comprising of architecture and procedure with the desirable features. In section 4, we give the experimental evaluation. In section 5, we give the concluding remarks and future research directions.

2. RELATED WORK

Expertise seeking is an information retrieval task concerned with the search for the most knowledgeable people in a specific research topic. This task involves taking a user’s query and returning a list of people sorted by their level of expertise regarding the query topic. In order to rank experts, we must evaluate them according to high-impact criteria termed features. They are extracted from different sources on the Web, for example from publications, one can extract the h-index, number of publications, number of citations...etc. It is also possible to consider other features collected outside publications, for instance a scholar’s awards, his courses...etc. All of these features are capable of indicating a scholar’s expertise but not with equivalent proportions.

Different models were proposed in this context; Generative, Voting, Graph-based and Learning to Rank models. Generative models rank candidate experts according to the likelihood of a person being an expert on the given query [1]. A researcher is evaluated through his documents, either through language models (LM) which look for the occurrence of query words, i.e. it uses terms to represent the content of publications or through topic models [[2], [3]] that detect the semantics of documents since it learns the topics found in the corpus with its related words through

unsupervised learning [[4], [5]]. In voting models, documents ranked with respect to a query vote for candidate experts in which they are mentioned [1]. Some voting methods are also applied in the aspect of rank aggregation [6]. Graph models are supplied with the ability of modeling various relations among people and documents, both implicit and explicit ones. Such relations are well represented by expertise graphs, where both documents and candidate experts become vertices and directed edges symbolize containment conditions. They can handle modeling document-candidate, document-document, and candidate-candidate associations [7]. Most of the previous models are not aware of handling heterogeneous features which probably needs a well-determined technique and learning to rank was there to fill such a prominent gap. Even though, voting models were somehow useful in the aforementioned task, but they only provided an acceptable result because regardless of data, the features are given the same weights. Learning to rank applies machine learning strategies to learn the ranking model from data in a supervised manner. In other terms, data is free to speak for itself rather than making assumptions regarding the model [8].

Some considered generative models in their work; H. Deng et al. [9] proposed a hybrid model by combining a weighted language model and a topic-based model. E. Smirnova et al. [10] proposed a user-oriented approach that balances two factors that influence the user's choice: time to contact an expert, and the knowledge value gained after. J. Tang et al. [11] proposed three generative topic models by introducing the conference information into the author topic model. Concerning voting models, D. Eddine Difallah et al. [12] developed a system capable of selecting which workers should perform a given task based on worker profiles extracted from social networks by considering the information extracted from the task descriptions and categories liked by the user on social platforms. Others applied graphs models; Zhou et al. [13] proposed a topic-sensitive probabilistic model, an extension of PageRank, which considers both the link structure and the topical similarity among users. J. Zhang et al. [14] introduced a propagation-based approach which estimates researchers' initial scores through their local information, then expertise of researchers is propagated towards the ones he co-authored with based on the intuition if a person knows many experts on a topic or his name co-occurs many times with another expert, then he is likely an expert. A. Kardan et al. [15] proposed a model for expert finding on social networks where people in social networks are represented instead of web pages in PageRank. Recent works have concentrated on learning to rank models; V. Kavitha et al. [16] combined multiple features of research expertise to rank experts; time weighted citation graph by giving significance to recent publications of an author and modified LDA to cope up with newly generated publication terms; Z. Yang et al. [17] used a supervised learning model with features including language models, author-conference-topic model, and other ones. C. Moreira et al. [18] proposed some features; academic indexes, regarding the textual content, Okapi BM25, TF and IDF were suggested in addition to some profile information; Sorg et al. [19] proposed a discriminative model that allows the combination of different sources of evidence in a single retrieval model using Multi-Layer Perceptron (MLP) and Logistic Regression as regression classifiers. The considered features were derived from language models, standard probabilistic retrieval functions and features quantifying the popularity of an expert in the question category.

2.1. Discussion

The previous works' analysis made us detect two key concepts which building an expert finding system relies on; the quality of the features incorporated into the system, i.e. their strengths of indicating the researchers' proficiency, and the way of exploiting them. As a matter of fact, the related work approaches were devoted towards the researcher's academic work, something he is

producing academically, i.e. publications, denoting his contributions and the topical domains he appears to be interested in. Yet, not only this kind of information it offers, but also one can infer the publications content by interpretation and analysis, their statistics and his relations. Apart from publications, a researcher tend to take part in activities revealing to some extent his proficiency level.

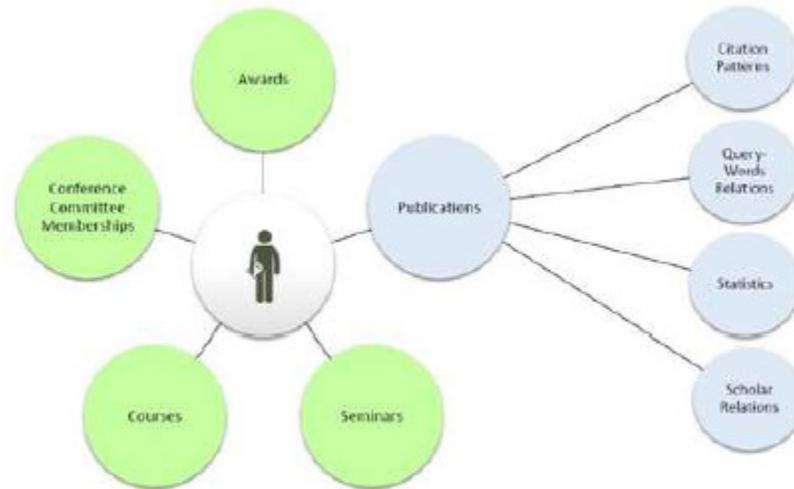


Figure 1. Scholar's Expertise Indicators

Figure 1 shows a scholar's expertise indicators. A scholar might be misjudged when evaluated only through his publications, for there are other unnoticed signs indicative of his expertise for example giving seminars, being part of conference committees, teaching courses and being honored and awarded. We aim at incorporating the conference memberships for it is important to study how good a scholar is at evaluating other scholars' academic work, where the classical case ranks him only based on his own academic work. Our proposed approach uses learning to rank for its ability of combining heterogeneous features optimally in a supervised manner; in addition to using the author topic model (generative model) to estimate the relation between author and topic through unsupervised learning. The reason why we picked this model is because: publications' abstracts are not always available which makes it pretty hard to know the topics of a publication when using LDA, i.e. authors' information are not included, and it links authors to topics in a semantic manner.

3. THE PROPOSED APPROACH

We aim at building an expert seeking system with the purpose of enhancing the retrieval effectiveness through adding non-traditional features by taking advantage of conference memberships following the intuition that such positions are not assigned spontaneously. Additionally, we believe that a scholars' publications conference rank, to some extent, matter; for experts usually publish their valuable work in top-ranked conferences. The more he publish in top-ranked conferences, the more valuable his work is. Our system is called "Multi-Feature Based Expert Seeking System" (FeBES) because it uses multiple features extracted from different sources to estimate scholars' expertise. In this section, the proposed model's concepts are provided, including the model's architecture, procedure and desirable features on which we will count to determine how professional scholars are given a domain.

3.1. Architecture

Figure 2 shows the architecture of the proposed approach, it is a learning to rank architecture. The approach is a mixture of models; it uses learning to rank to combine multiple features, and uses generative model (author topic model) to estimate the relation between author and query through unsupervised learning to link topics with document words and authors.

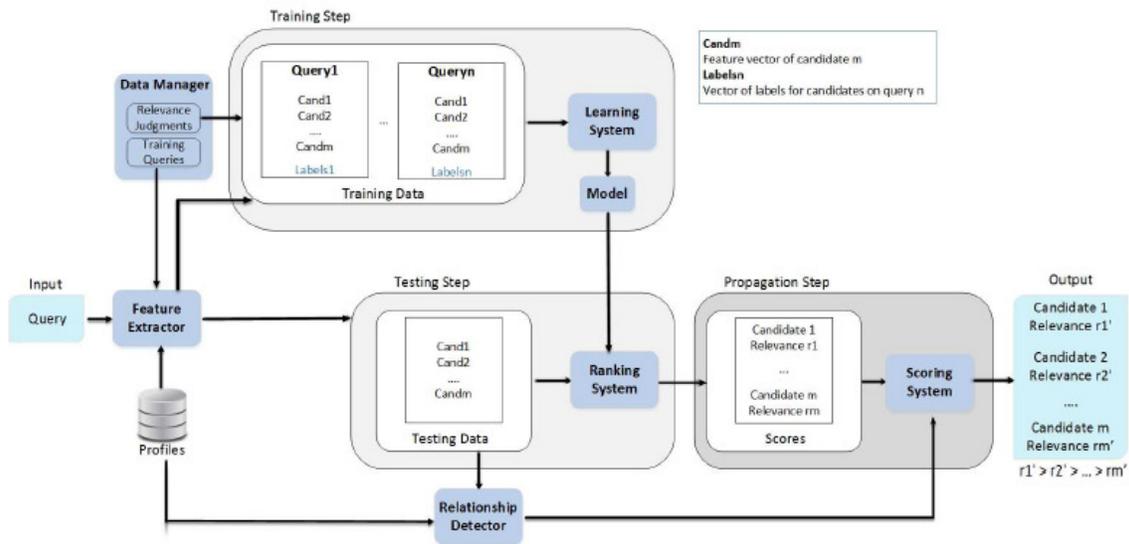


Figure 2. Our Approach Architecture

3.2. Procedure

The input represents the topic of interest submitted by the user. The set of profiles belong to researchers where each profile contains personal information and the academic work done by the researcher, having the following attributes: name, affiliation, e-mail, location, homepage, summary about his career and a list of all his publications. These profiles are generated in [20] through correlating the information extracted from heterogeneous sources, by taking advantage of data repetition in multiple sources and those existing in one source. On one hand, data is being validated and links between sources are created. On the other hand, the missing data issue is solved, and are saved in the repository serving as the system's internal input. The data manager is in charge of maintaining the training queries with relevance judgments of researchers with respect to those queries. Features are extracted based on the researchers' profiles by the feature extractor, they are of two types; query-independent and query-dependent. The input into the learning system is the training instances. Each training instance refers to a query associated with feature vectors resulting from all features for researchers along with their relevance judgments with respect to that query. This component generates the learned ranking model, comprising of features' weights, which serves as an input to the ranking system to handle all user-submitted queries. The ranking system receives the user's input query, then retrieves the candidate experts from the set of all researchers. Afterwards, feature vectors for those candidates are built. Finally, the learned model is applied on the constructed feature vectors to generate a score for each candidate. The output displayed to the user is a ranked list of candidate experts by sorting the scores generated by the ranking system.

3.3. Features

Features are of two types: query-independent and query-dependent. Features belonging to the former category maintain the same value regardless the query topic, however those belonging to the latter one changes as the query changes. We will mention the included features and how to estimate their values.

Query-Independent Category

We have cautiously picked the features which seem to mostly influence a researcher's expertise. The considered features are; h-index, i10-index, total number of citations, total number of publications, number of publications in recent years, total number of memberships in conferences, and conferences' rank. We believe experts are more likely to publish papers in high-ranked conferences. It can be estimated in the following way:

$$\left(\sum_{\text{rank}} W_{\text{rank}} * n_{\text{pubs}(\text{rank})} \right) / n_t \quad (1)$$

Where rank represents a conference rank and wrank is the weight of this rank. Parameter npubs(rank) is the number of papers published in conferences having this rank. The overall score is divided by the total number of publications of the scholar n_t because we are into knowing the distribution of all publications over these ranks.

Query-Dependent Category

We relied on the relation between query and author $P(q, a)$ through Author-Topic model. The more he has published on-topic publications, the stronger $P(q, a)$ is, because as a result, he would be assigned more to on-topic words. The following equation was applied

$$P(q, a) = \sum_{w_i} \sum_t P(w_i|t) P(a|t) P(t) \quad (2)$$

Where w_i is a query word, t is a topic among the set of all topics. $P(w_i|t)$ is the probability of word w_i given topic t , $P(a|t)$ is the probability of author a given topic t and $P(t)$ is the prior probability of topic t .

We also considered the scholar's conference memberships. The intuition is that scholars who very often take part as members in conferences should be thought of as more valuable due to the fact such positions are not assigned spontaneously. Our concerns include how often he is a part of conference committees, how connected these conferences are to the query terms and how dominant these relevant participations are, relative to all other ones. Though the latter point may sound odd, most experts are fully devoted to one domain and partially to related ones. We also counted on the conference rank based on its importance. The following formula is inspired from the document model, but it has been applied only on documents, we projected it to our context, where documents were replaced by conferences and we introduced another value which influences the whole formula.

$$P(q|a) = \alpha \sum_c P(q|c) P(a|c) P(c) \quad (3)$$

Where q and a represent the query and author respectively, and c is a conference. $P(q|a)$ is the probability of query q given author a , $P(a|c)$ is the probability of author a given conference c , $P(q|c)$ is the probability of query q given conference c and $P(c)$ is the prior probability of conference c . $P(a|c)$ is either zero or one depending on whether or not author a is in the committee of conference c . $P(c)$ depends on the conference c rank. To estimate $P(q|c)$, conference c needs to be represented as the collection of documents that published in conference c with the use of language and author topic models.

$$P(q|c) = \sum_d P(q|d) = \sum_d \prod_{w_i} P(w_i|d) = \sum_d \prod_{w_i} P_{LM}(w_i|d) \times P_{AT}(w_i|d) \quad (4)$$

$$P_{AT}(w|d) = \sum_t \sum_a P(w|z) P(z|a) P(a|d) \quad (5)$$

Where t is a topic and a is an author of document d . regarding the author topic model. $P(w|z)$ is the probability of word w given topic z , $P(z|a)$ is the probability of topic z given author a and $P(a|d)$ defines the relation between author a and document d .

$$P_{LM}(w|d) = (1 - \lambda) P(w|d) + \lambda P(w) \quad (6)$$

Regarding the language model, $P(w|d)$ is the probability of word t in document d , $P(w)$ is the probability of t in the whole corpus and λ is a smoothing parameter. As for α , it represents the dominance degree of on-topic memberships. To apply this, we have to estimate $P(q|c)$ for each conference and decide, based on a threshold value, whether or not conference c is highly related to query q .

$$\alpha = n/n' \quad (7)$$

where n is the number of strongly on-topic participations and n' is the number of all participations including relevant and non-relevant ones.

4. EXPERIMENTATION & EVALUATION

4.1. Implementation & Preprocessing

We have implemented the various system elements, and thus provided web interface for receiving user requests and respond with relevant results. The prototype of our architecture is implemented using JavaEE, where all the tests were performed on 2.60 GHz Processor, 8GB of RAM PC. Additionally, we used Matlab to apply “author-topic model”, which is an unsupervised learning technique. Our system is portable and domain-independent, i.e. it can be applied on any dataset regardless of its domain. To test our approach, we have chosen a Computer Science dataset with 297 researchers including around 215 experts from 7 topics.

With regard to author topic model, we considered 54,000 publications. The dataset has undergone some preprocessing (insignificant authors removal, associated publications removal, stop words and insignificant words removal) and the number of topics was set to 80 and number of iterations to 1000. We were left with 8,534 authors, 31,114 publications, 19,607 vocabulary words and 1,640,069 word tokens. The output is a list of learned topics having topic-author and topic-word proportions

We used SVMRank to learn the model and considered three queries. We used the Arnetminer's evaluation dataset comprising of seven queries. For each query, we labeled experts list with 1, and complemented it with equivalent number of non-experts (labeled as 0), containing easy and hard examples. Researchers' feature vectors undergone some preprocessing as well by removing the null and conflicted ones and finally normalizing them.

4.2. Results & Evaluation

Table 1 shows the learned model comprising of the features' weights. The obtained results reveal that conference committee memberships has the biggest impact among all other features, for experts are members of committees just for they have the required expertise. Another high influential information is the h-index, because it represents the scientific impact of his publications on other scholars, it also signifies his productivity. The relation between query and author through author-topic model is significant as well, because this value is based on the author's publications, the more he is assigned to query topic-related words in his publications, the more he is likely one of the experts on the query topic. As for conference ranks, the other introduced criterion, showed that it does influence the scholar's expertise but not in same degrees as the previous ones. Even the number of publications has not a very big impact, because it's not always about the quantity of publications, it's more about the quality. The weight for number of citations has a negative impact, even though the h-index had the opposite case. Well, this is because two scholars might have the same total number of citations, but the way they are distributed over their publications may vary (number of publications may vary too) and when that happens, the h-index varies as well. The number of publications in recent years (2009 and above) showed that it's not important for a person to have published a lot of papers in the recent 6 years, the whole career counts more than these last years. I10-index's weight is the smallest one, and it is that way perhaps because h-index almost contains the i10-index concept, and the former highly influences the expertise.

Table 1. Features Weights

Feature	Weight
# of pubs	0.810033
# of recent pubs	-1.703149
# of citations	-0.440943
H-index	3.464950
I10-index	-1.323053
Conference ranks	1.118827
Author-topic model	2.461189
Conference committee	4.592299

For the evaluation, we used Arnetminer's new dataset (7 queries). Precision at 5, 10, 15, 20 (P@5, P@10, P@15, P@20) and the Mean Average Precision (MAP) are the considered evaluation metrics. Tests were done on two approaches, a baseline approach and our approach where our approach = baseline + {conference rank, conference committee}. The figures 2, 3 and 4 below show the performance of the baseline approach and our approach in Information Extraction, Machine Learning and Intelligent Agents respectively.

In figure 3, we can clearly see the gap between the two approaches. When n was 5, the baseline precision was very low (20%), whereas that of our approach was 60%. Even though when n

became 10, the precision of the former improved and that of the latter declined, but the latter still had a better precision. Later on, both precisions dropped with our approach outperforming the baseline approach. In figure 4, we can notice that baseline approach and our approach have the same initial value at $p@5$, but when n increased to 10, the baseline precision decreased, however that of our approach maintained the same value. Until n became 20, our approach outperformed the baseline approach. In figure 5, both approaches had the same initial precision value. And when the approaches' precision began to reduce, our approach still outperformed the baseline approach.

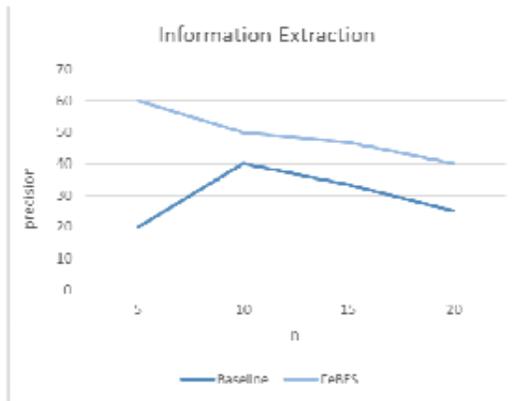


Figure 3. Comparison between Baseline and Our Approach in Information Extraction

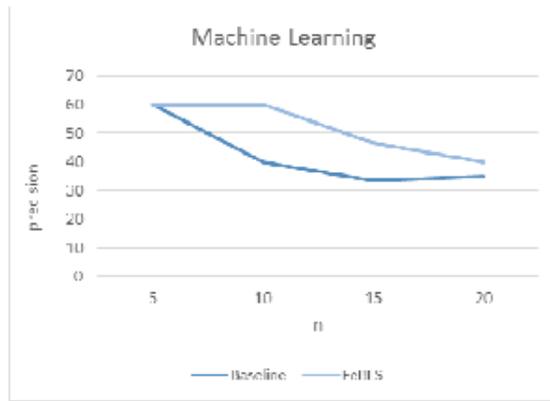


Figure 4. Comparison between Baseline and Our Approach in Machine Learning

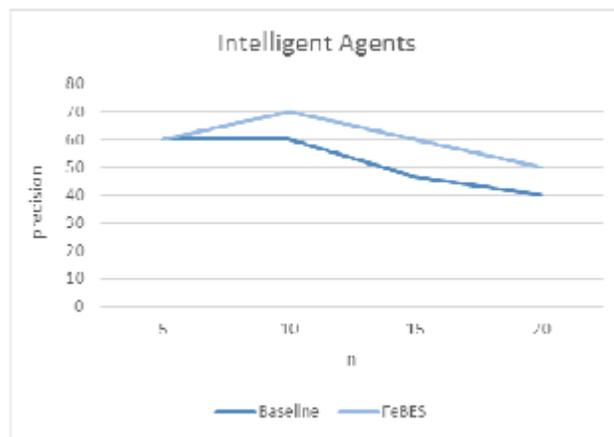


Figure 5. Comparison between Baseline and Our Approach on Intelligent Agents

Table 2 summarizes the three queries (by taking the average), and estimates MAP for both approaches. We can easily tell the difference between baseline approach and our approach. At the four n values, the latter had a better performance and a more effective result than that of the former. When n was 5, 10 and 15, the difference was around 14%. As for the MAP, the results showed our approach outperformed the baseline by 18%. The explanation behind such results is because publications don't cover the same evidence as that of conference memberships, where the latter helps distinguish the significant researchers according to the topic they are experienced in

based on being a member in conferences. In addition to the papers' conference ranks, where the quality of a scholar's papers is considered, so it is not only about the quantity of these papers but also their quality and value.

Table 2. Comparison between Baseline and Our Approach

	P @ 5	P @ 10	P @ 15	P @ 20	MAP
Baseline Approach	46.7%	46.7%	37.9%	33.4%	39.1%
Our Approach	60%	60%	51.2%	43.4%	57.5%

5. CONCLUSION & FUTURE WORK

The previous results show that our approach has 57.5% as mean average precision whereas 39.1% for the baseline. Concerning the three queries, our approach outperforms the baseline. Based on these results, we noticed that including new kind of expertise evidence into expert seeking came in handy because they happen to represent a different aspect of expertise. Therefore, we conclude that a scholar's expertise should be assessed not only through his academic publications, but also through external activities he is involved in because ignoring such an aspect might cause misjudgment.

As future plans, we aim at verifying the effectiveness of other beyond publication features including awards, courses, and seminars. Not to forget to mention that applying the propagation phase, known as propagating one's expertise based on those he has relations with, has shown enhancement as stated in the related work section. For this reason, we intend to do a combination by adding the propagation phase to our approach to improve the retrieval mperformance. We also would like to prove the efficiency of orienting this phase towards time and distinguishing between co-author and supervision relations. Moreover, we believe it is preferable to distinguish between the different roles a scholar is assigned when he is in a conference committee, because he could possibly be in a scientific committee, or in a program committee or even a reviewer.

REFERENCES

- [1] K. Balog, Y. Fang, M. d. Rijke, P. Serdyukov and L. Si, Expertise Retrieval, Foundations and Trends in Information Retrieval, Vol. 6: No. 2–3, pp 127-256., 2012.
- [2] C. Moreira, "Learning to rank academic experts," Master Thesis, 2011.
- [3] C. Zhai, "Statistical Language Models for Information Retrieval A Critical Review," in Foundations and Trends in Information Retrieval, vol. 2, 2008, p. 137–213.
- [4] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth and M. Steyvers, "Learning Author- Topic Models from Text Corpora," in ACM Transactions on Information Systems (TOIS), 2010.
- [5] K. Balog and M. d. Rijke, "Associating people and documents," in Proceedings of the European Conference on IR Research, (ECIR '08), Berlin, Heidelberg, 2008.
- [6] C. Moreira, P. Calado and B. Martins, "Learning to Rank Experts in Academic Digital Libraries," in 15th Portuguese Conference on Artificial Intelligence, EPIA, Lisbon, Portugal, 2011.

- [7] P. Serdyukov, H. Rode and D. Hiemstra, "Modeling multi-step relevance propagation for expert finding," in Proceeding of the ACM International Conference on Information and Knowledge Management, New York, NY, USA, 2008.
- [8] C. Moreira, "Learning to rank academic experts," Master Thesis, 2011.
- [9] H. Deng, I. King and M. R. Lyu, "Formal models for expert finding on DBLP bibliography data," in Proceedings of the IEEE International Conference on Data Mining, Washington, DC, USA, , 2008.
- [10] E. Smirnova and K. Balog, "A User-Oriented Model for Expert Finding," in LNCS 6611, 2011.
- [11] J. Tang, J. Zhang, J. L. L. Yao, L. Zhang and Z. Su, "Arnetminer: Extraction and mining of academic social network," in Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD '08, New York, NY, USA, 2008.
- [12] D. E. Difallah, G. Demartini and a. P. Cudré-Mauroux, "Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do, A Crowdsourcing Platform for Personalized Human Intelligence Task Assignment Based on Social Networks," in WWW, 2013.
- [13] G. Zhou, S. Lai, K. Liu and J. Zhao, "Topic-Sensitive Probabilistic Model for Expert Finding in Question Answer Communities," in CIKM, Maui, HI, USA, 2012.
- [14] J. Zhang, J. Tang and a. J. Li, "Expert Finding in a Social Network," in 12th International Conference on Database Systems for Advanced Applications, Bangkok, Thailand, 2007.
- [15] A. Kardan, A. Omidvar and Farahmandnia, "Expert finding on social network with link analysis approach," in Electrical Engineering (ICEE), 19th Iranian Conference, 2011.
- [16] V. Kavitha, G. Manju and T. Geetha, "Learning to Rank Experts using combination of Multiple Features of Expertise," in Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference, 2014.
- [17] M. n. uddin, t. h. duong, K.-j. oh, j.-g. jung and g.-s. jo, "experts search and rank with social network: an ontology-based approach," International Journal of Software Engineering and Knowledge Engineering, vol. 23, no. 1, 2013.
- [18] M. Rosen-Zvi, T. Griffiths, M. Steyvere and P. Smyth., "The author-topic model for authors and documents," in UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence.
- [19] P. Sorg and P. Cimiano, "Finding the Right Expert: Discriminative Models for Expert Retrieval, in Proceedings of the International," in Conference on Knowledge Discovery and Information Retrieval (KDIR), Paris, France, 2011.
- [20] H. Nouredine, I. Jarkass, H. Hazimeh, O. A. Khaled and E. Mugellini, " CARP: Correlationbased Approach for Researcher Profiling," in 27th International Conference on Software Engineering and Knowledge Engineering SEKE, 2015.

AUTHORS

Mariam Abdullah is a Software Developer at Mentis Company in Beirut, Lebanon. She was born in Tyre, Lebanon in 1992. She received her Bachelor's degree in Computer Science in 2013 from the Lebanese University, Beirut, Lebanon. She received her Master's degree in Information and Decision Support Systems in 2015 from the Lebanese University in collaboration with university of Applied Sciences of Western Switzerland (HES-SO).



Hassan Nouredine received the Ph.D. degree from the Lebanese University in cooperation with the University Of Applied Sciences Of Western Switzerland (HESSO-FR), Fribourg in 2016. He received Research Master Degree in Signal, Telecom, Image, and Speech (STIP) from the Lebanese University, Doctoral School of Sciences and Technology in 2011. Since 2015, he has been with the American university of education and culture where he is lecturing in Computer Science Studies at the Faculty of Science. His main areas of research interest are Semantic Web & Ontologies, Social Network Analysis, Sentiment analysis, Identity linkage and Information extraction.



Jawad Makki received the Ph.D. degree from the University of Toulouse, France, in 2010. Since 2011, he has been with the Lebanese University where he is lecturing in Computer Science and Information Studies at the Faculty of Science and the Faculty of Information. His main areas of research interest are Semantic Web & Ontologies, Knowledge Representation, NLP & Information Extraction, and Social Network Analysis.



Hussein CHARARA has received an M.S. degree in Computer and Communications Engineering from the Lebanese University, Lebanon in 2002, and an M.S. degree in Networking and Telecommunications from the Institut National Polytechnique (INP-ENSEEIH), France, in 2003. In 2007, He obtained a Ph.D. degree in Network, Telecom, Systems and Architecture from INP - IRIT, France. From 2006 to 2009 He worked for AIRBUS and THALES AV avionics as a R&D Engineer and PM. He contributed to the implementation & development of the Real time embedded AFDX networks such as for A380, A400M and Soukhoï RRJ programs. From 2009 to 2010, He worked for the ARECS GmbH - München where He was involved in the GPS/Galileo receivers modeling and simulations. Since 2010, He is an assistant professor at the Lebanese University. He is working in several research fields including: Traffic engineering of real time network, Performance Evaluation and QoS, Avionics and wireless sensor networks



Hussein Hazimeh is a PhD student at the University of Fribourg, Fribourg, Switzerland under the supervision of Professor Philippe Cudré-Mauroux, in collaboration with the university of Applied Sciences of Western Switzerland (HES-SO) under the supervision of Professor Omar Abou Khaled and Professor Elena Mugellini. He received his Master 2 from the Lebanese University, Beirut Lebanon, in collaboration with HES-SO. His research interests focus on: Social Media Analytics, Sentiment analysis, Identity linkage and Information extraction.



Omar Abou Khaled is Professor in computer science at HES-SO campus Fribourg (EIA-FR) and member of HumanTech Institute (humantech.eiafr.ch/). He holds a PhD in Computer Science received from the Perception and Automatic Control Group of HEUDIASYC Laboratory of "Université de Technologie de Compiègne", and a Master in Computer Science from the same university. Since 1996 he has been working as research assistant in the MEDIA group of the Theoretical Computer Science Laboratory of EPFL (Ecole Polytechnique Fédérale de Lausanne) in the field of Educational Technologies and Web Based Training research field on MEDIT and CLASSROOM 2000 projects.



He was International Advisor at HES-SO until August 2014. He was Head of Internationale Relations Office at EIA-FR. Head of MRU "Advanced IT Systems Architects" at EIA-FR. Until 2007 He was leader of the MISG (Multimedia Information System Group). He is responsible of several projects in the field of Document Engineering, Multimodal Interfaces, Context Awareness, Ambient Intelligence, Blended Learning, and Content-Based Multimedia Retrieval. He is involved in the teaching of several courses related to Information Systems, Web Technologies & Multimodal Interfaces.

Elena Mugellini is currently Professor at the University of Applied Sciences of Western Switzerland in Fribourg (HES-SO). She holds a Diploma (Bsc and Msc) in Telecommunications Engineering and a Ph.D. in Computer Sciences from University of Florence, Italy. Elena is the leader of HumanTech Institute (Technology for Human well-being, humantech.eia-fr.ch/). Her research expertise lies in Human Computer Interaction (natural interaction, smart spaces, machine learning, serious game and gamification) and Intelligent Data Analysis (multimedia content and knowledge management, semantic technologies, information visualization).



She teaches at graduate and undergraduate level a range of topics, including: project management, human-computer interaction and user experience design, web engineering, software engineering and information systems. Her work is published in journals, conference proceedings and as book chapters and has also been presented in numerous scientific conferences. She is an active reviewer for several international conferences and journals.

INTENTIONAL BLANK

ESTIMATING HANDLING TIME OF SOFTWARE DEFECTS

George Kour¹, Shaul Strachan² and Raz Regev²

¹Hewlett Packard Labs, Guthwirth Park, Technion, Israel

²Hewlett Packard Enterprise, Yehud, Israel

ABSTRACT

The problem of accurately predicting handling time for software defects is of great practical importance. However, it is difficult to suggest a practical generic algorithm for such estimates, due in part to the limited information available when opening a defect and the lack of a uniform standard for defect structure. We suggest an algorithm to address these challenges that is implementable over different defect management tools. Our algorithm uses machine learning regression techniques to predict the handling time of defects based on past behaviour of similar defects. The algorithm relies only on a minimal set of assumptions about the structure of the input data. We show how an implementation of this algorithm predicts defect handling time with promising accuracy results.

KEYWORDS

Defects, Bug-fixing time, Effort estimation, Software maintenance, Defect prediction, Data mining

1. INTRODUCTION

It is estimated that between 50% and 80% of the total cost of a software system is spent on fixing defects [1]. Therefore, the ability to accurately estimate the time and effort needed to repair a defect has a profound effect on the reliability, quality and planning of products [2]. There are two methods commonly used to estimate the time needed to fix a defect, the first is manual analysis by a developer and the second is a simple averaging over previously resolved defects. However, while the first method does not scale well for a large number of defects and is subjective, the second method is inaccurate due to over-simplification.

Application Lifecycle Management (ALM) tools are used to manage the lifecycle of application development. Our algorithm relies on a minimal number of implementation details specific to any tool and therefore has general relevance. Our implementation is based on the *Hewlett Packard Enterprise* (HPE) ALM tool. We tested the algorithm with projects from different verticals to verify its broad applicability.

One of the advantages of our implementation is that it does not assume a standard data model, but is able to handle the cases where the defect fields available vary between different data sets. We

provide experimental evidence of the significance of including non-standard fields as input features for the learning algorithm.

Our approach supposes that we can learn from historical data on defects that have similar characteristics to a new defect and use this to make predictions on the defect handling time that are more accurate than many comparative approaches. We found that in real life, the available data was often of poor quality or incomplete, and so in implementing our solution we encountered a number of challenges, which we discuss in this paper.

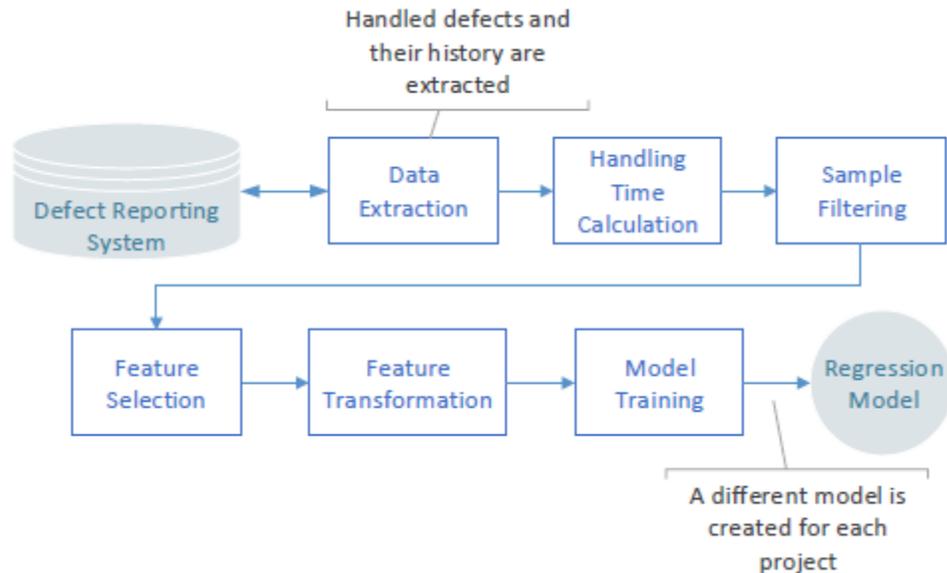


Figure 1: Training process

Our approach is based on a supervised regression machine learning algorithm in which the output is the predicted handling time for defects. The training phase is summarized in Figure 1. It takes as its input (1) a training set of handled defects together with their fields and (2) the history of status changes from these defects, and outputs a learned model. The prediction phase is summarized in Figure 2. It takes as its input (1) the model outputted from the training phase and (2) a set of unhandled defects together with their fields, and outputs the predicted handling time.

2. RELATED WORK

There are many practical methods of defect handling time estimation used in software engineering, as described in [2].

Defect fix effort estimation using neural networks was reported in [3]. The data for this study was extracted from the NASA IV&V Facility Metrics Data Program (MDP) data repository. Their approach is based on clustering the defects using a Kohonen network. Then the known values of defects fix effort were assigned to the found clusters. Given an unseen sample, the fix time is estimated based on the probability distribution of the different clusters.

A text similarity approach for estimating defect resolution time was described in [4]. The title and description of the defects were used to measure similarity between defects using the "Lucene"

engine developed by the Apache foundation [5]. In this study, *k Nearest Neighbours* (k-NN) was used to find the closest *k* defects already resolved to a given new defect, and the mean was taken as the final estimation.

A work done by Rattiken and Kijsanayothin in [6] investigated several classification algorithms for solving the defect repair time estimation problem. The data set, taken from defect reports during release of a medical record system, contained 1460 defects. They investigated seven representative classification variants of the following algorithm families: Decision Tree Learner, Naive Bayes Classifier, Neural Networks (NN), kernel-based Support Vector Machine (SVM), Decision Table, and *k*-NN.

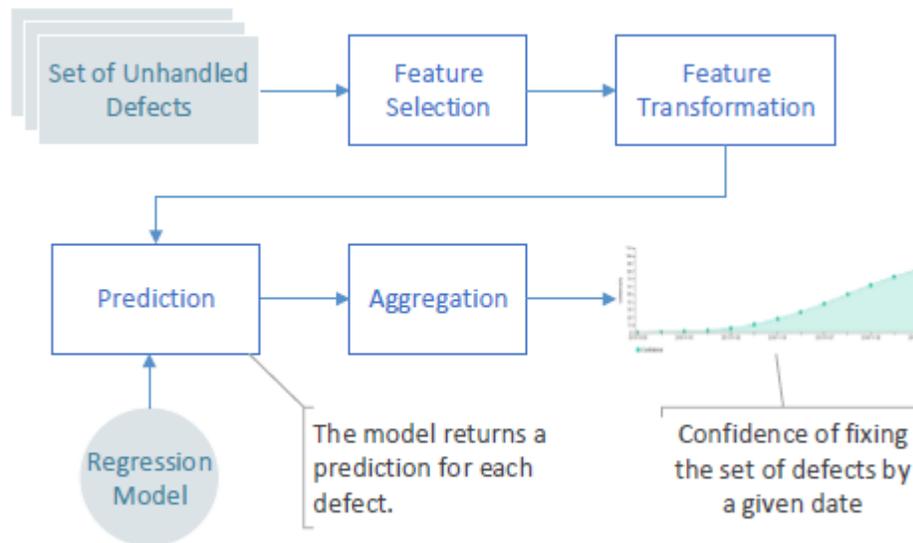


Figure 2: Prediction process

Giger et al. presented a method for predicting the handling time of a new defect using decision trees [7]. The defects in the sample set were classified into two categories: fast and slow defect fixes, based on the median defect fix time of the corresponding project. The data set included 6 open source projects, each containing between 3,500 and 13,500 defects spanning periods between 6 and 11 years.

In a recent work by Zhang et al. in [8], the authors suggested a k-NN classifier for predicting fast or slow defect fixes. A bug was represented as a vector of standard defect attributes and a predefined bag of words extracted from the summary field. They used defect datasets collected from three CA Technologies projects. The results reported in this work is based on a normalized time unit, where 1 unit equals the median time needed to fix a defect.

3. TRAINING

3.1. Data Extraction and Handling Time Calculation

In the first step in the training process, we first extract the defect records, including all their fields and history from the project database. The fields available vary from project to project. Typical

fields include Summary, Description, Assignee, Status, and Severity. However, users generally customize their projects to include user-defined fields. Excluding such fields from the analysis might mean missing valuable features. For example, suppose there is a field such as Estimated Effort, with possible values Low, Medium and High. It seems reasonable that a correlation exists between this field and the actual handling time. Our system does not rely on knowing in advance which fields exist and is able to learn from any number of fields, of various types. We use the history of the Status field to calculate the total time effort invested in fixing a handled defect, referred to as Handling Time. Possible defect statuses vary between projects but generally include New, Open, Fixed, and Closed. For the purposes of our research we define our target value as the total number of days spent in status Open. For the training set, we consider only defects that have reached the end of their lifecycle, i.e. are in state Closed.

In Table 1 we see general information about the projects used in this study. The data sets used for analysis were snapshots of customers' databases.

Table 1: Summary of the projects in the data set

Project	Total # of defects	Observation Period	Industry
1	41,490	Jan. 2010 - Dec. 2013	Banking
2	15,291	Nov. 2004 - May. 2007	Banking
3	3,851	Sep. 2012 - Nov. 2015	Telecom.
4	48,855	Mar. 2001 - Oct. 2006	Software
5	29,425	Sep. 2001 - Dec. 2006	Software
6	2,350	Oct. 2013 - Nov. 2015	Software

3.2. Sample Filtering

We aim to provide managers with a tool to enable better estimation of the time needed to handle defects. Therefore we cannot allow the prediction system to estimate a practically unreasonable time for fixing a defect and we consider it as sampling error. Such a long fixing time can be caused by the fact that in some cases users do not update their current status on working on a defect, and may mark a defect as Open, and after a while switch to another task without updating the defect status.

Although other related work in the field allowed long fixing periods (e.g. more than a month) [7], we observed that it is rare that a defect takes more than 30 days to be fixed and so such defects were filtered out. We encountered defects that remained in state Open for more than a year. Figure 3 shows the accumulative distribution of the defect handling time versus their percentage in the data set. If we allow defects with very large fixing time, the algorithm might find patterns characterizing defects that were "forgotten" in Open state rather than identifying patterns that affect real-life handling time.

In addition, we employ standard extreme value analysis to determine the statistical tails of the underlying distribution using the z-scoring on the handling time field [9]. Samples that are not in the interval $[\mu - 3\sigma, \mu + 3\sigma]$ are filtered out, where μ is the mean and σ is the standard deviation of all the defects in the data set.

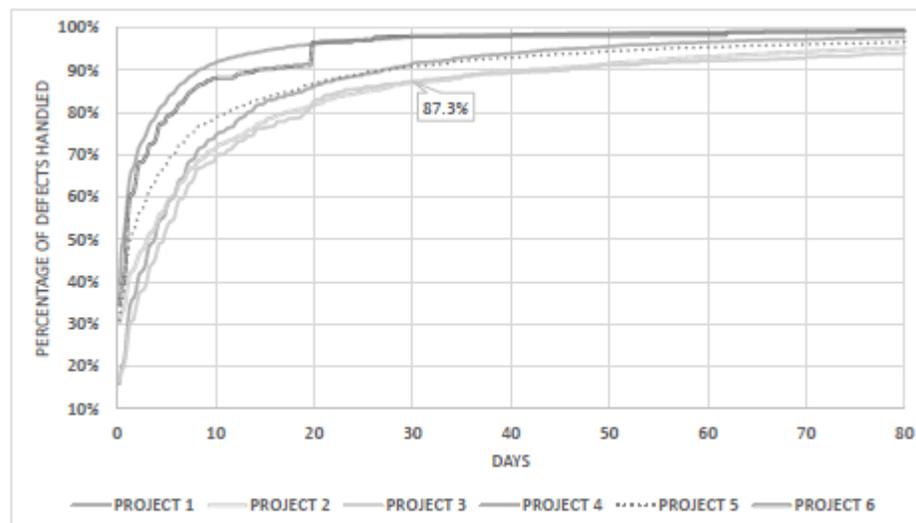


Figure 3: Percentage of defects handled by days

3.3. Feature Selection

Different fields in a data set may have different population rates, where the population rate is the percentage of samples containing a value in a given field. This might be due to the fact that not all fields are required for each defect, or that a field was added to the project at a later stage, and therefore defects opened prior to this field's addition do not have values for the field. To improve the data set stability and the quality of the model, we remove fields with very small population rate (less than 2%). To avoid data leakage, the system automatically filters out attributes that are usually unavailable on defect submission, by comparing the population rate of the attributes in handled and unresolved defects. We also remove non-relevant internal system fields.

3.4. Feature Transformation

A defect is represented by a set of data fields of both qualitative and quantitative data types. However, most machine learning algorithms employ quantitative models in which all variables must have continuous or discrete numerical values. Therefore, certain transformations must sometimes be performed on the given fields before they can be used as features to the learning algorithm.

Numeric fields are generally used as is, directly as features. However, in special cases they are treated as categorical features, as described later in this section.

Categorical fields are transformed using a "one-hot" encoding scheme; a vector of features is assigned to each such field - each feature corresponding to a possible value of that field. For each defect, the feature corresponding to the actual value of the field in that particular defect is assigned the value 1, whereas the rest of the features are given the value 0. We identify a categorical field as a string or number field whose number of distinct values is sufficiently lower than the number of defects with this field populated. To avoid there being too many features ("the curse of dimensionality" [10]) we group the values with low frequencies together under a single value Other, which also captures possible unprecedented values during the prediction process.

String fields whose domain of values is very large, may correspond to free-text fields such as Summary and Description. Although these fields may contain important information, they require additional effort to be appropriately mined and so were discarded in the current implementation.

Date and time fields are projected into time/date cycles such as hour of the week, used to capture the presumed correlation between the day of week for some fields and the handling time. In order to represent this feature radially, we projected each date onto a unit-circle representing the corresponding hour of the week h and took $\cos(h)$ and $\text{sign}(\sin(h))$, where $\text{sign}(\cdot)$ is the signum function, as a new pair of independent variables.

Scarcity of data is a common issue when analysing real-life project defects. Therefore, before starting the training phase, imputation is applied to fill missing values. Empty cells in categorical features are simply treated as another category, as they may have a particular meaning in many cases. However, for numeric features, we filled up empty cells with the most frequent value. In order to avoid data leakage, the imputation is performed just before training the model (after the partitioning to test and train). Replacing empty cells with the mean value of the feature or the median value are two other common numerical imputation schemes which were considered but were empirically found to yield inferior scores in our setting.

3.5. Model Training

Unlike in classification, in which the goal is to identify to which class an unlabelled observation belongs, regression algorithms, given a predictor variable x , and continuous response variable y , try to understand the relationship between x and y , and consequently enable predicting the value of y for a new value of x . The basic idea underlying the Regression Tree learning algorithm is similar to the idea on which the commonly used Decision Tree algorithm is based, but slightly altered to adapt the non-discrete nature of the target field. Random Forests [11] are an ensemble learning method for both classification and regression problems. They operate by building a multitude of decision trees and returning the class that is voted by the majority of the trees in classification, or the mean of the individual trees in regression [10]. Random forests are much less prone to overfitting to their training set compared to decision trees, and so we chose to use them as our machine learning algorithm.

We trained a random forest regression in Python using 80 estimators (i.e. individual decision trees), with a limitation on the minimum number of samples required to split internal node set to 2, and minimum number of samples in a newly created leaf set to 6. The default values were used for the other model parameters. These parameters are constant for all data sets in this work and they were empirically tuned.

4. PREDICTION

While estimating the handling time of a single defect is important, in reality managers are usually interested in making a prediction based on a set of defects (such as the content for a future release) to be handled by a group of developers. Therefore, we have built a system that provides a completion time estimation for any given set of unhandled defects.

As shown in Figure 4, the system presents a graph showing the resulting cumulative distribution in a report which displays the confidence of the defects being closed by any selected date. Based

on the graph, we let the user calculate the release end date given confidence level and vice versa. We also let the user set the number of available developers.

Figure 2 describes the flow for predicting the completion time for a set of defects. Each defect in the given set of unhandled defects passes through the same preprocessing flow as in the learning process, apart from the handling time calculation stage. Then, using the model, the system returns an estimation of the handling time of the given defect, as well as a prediction confidence calculated based on the variance between the answers of the individual regression trees.

To automatically calculate the total time it takes to complete a set of defects by several developers, one would ideally use the optimal scheduling which minimizes the makespan (i.e. the total length of the schedule). However, the problem of finding the optimal makespan in this setup, better known as the Minimum Makespan Scheduling on Identical Machines, is known to be NP-Hard [12]. We employ a polynomial-time approximation scheme (PTAS) called List Scheduling, utilizing the Longest Processing Time rule (LPT). We start by sorting the defects by non-increasing processing time estimation and then iteratively assign the next defect in the list to a developer with current smallest load.

An approximation algorithm for a minimization problem is said to have a performance guarantee p , if it always delivers a solution with objective function value at most p times the optimum value. A tight bound on the performance guarantee of any PTAS for this problem in the deterministic case, was proved by Kawaguchi and Kyan, to be $\frac{1+\sqrt{2}}{2}$ [13]. Graham proved a relatively satisfying performance guarantee of $4/3$ for LPT [14].

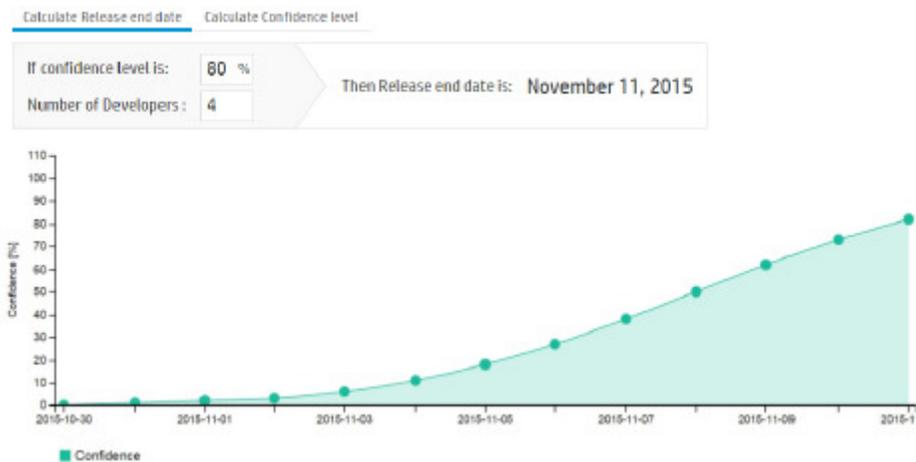


Figure 4: A screenshot of the actual system report.

After computing the scheduling scheme, we end up with a list of handling times corresponding to developers. The desired makespan is the maximum of these times.

Accounting for the codependency between defects is infeasible since it requires quantification of factors such as developers' expertise levels and current availability. It is also subject to significant variance even within a given project. In practice, we observed that treating defects as independent for the sake of this calculation yields very reasonable estimates. By the assumption that defects

are independent, we may treat the completion times as pairwise independent, which yields the Cumulative Distribution Function (CDF) $F(t)$. This value is the probability that the set of defects will be completed until time t .

$$F(t) = Prob(\max_i C_i \leq t) = \prod_i Prob(C_i \leq t) \quad (1)$$

Where C_i is the completion time for developer i . C_i 's distribution is not obvious. It is the sum of several independent random variables - a variable whose Probability Density Function (PDF) can be generally computed by convoluting the PDFs of its constituents. However, this requires knowing the distribution of each defect's handling time.

Zhang et al [8] found that the distributions best describing defect handling times are skewed distributions, such as the Weibull and Lognormal distributions. Therefore, we at first used to take the mean and variance values outputted by the predictor, fit a Weibull distribution corresponding to these values, and then apply convolution to achieve the required measures. Our results showed that, in most cases, the distributions of single defects highly resembled the Normal distribution. Moreover, convoluting their PDFs proved to converge very quickly to the Normal distribution, as the Central Limit Theorem guarantees. For these reasons, and to allow fast computation, we simplified the aggregation such that each variable was treated as a Normal random variable. The sum of such variables is also normally distributed and can be easily calculated by a closed formula.

Given $F(t)$, the expected value and variance of the total handling time can be derived using standard techniques.

5. EXPERIMENTAL METHODOLOGY

To facilitate comparison with related work, which mostly discuss handling time for individual defects, and due to the more accurate historical information available for such cases, we focused on the prediction accuracy for handling a single defect. To evaluate the quality of our predictive model, we use the six customer projects introduced in Table 1. After extracting a data set of defects and applying the preprocessing described in Section , we randomly partition the sample set into learning and testing sets, containing 80% and 20% of the data respectively.

The total number of defects in the sample sets of each project is shown in Table 2. The training set is used to construct the model. We use the standard accuracy estimation method, n-fold cross-validation [15] to avoid overfitting.

We employ several performance metrics to evaluate our algorithm. First, we use the *Coefficient of Determination* (denoted R^2) which is a key indicator of regression analysis. Given a set of observations $\{y_i\}_{i=1}^n$ with average value \bar{y} , in which each item y_i corresponds to a prediction p_i , R^2 is defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - p_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (2)$$

An R^2 of 1 indicates that the prediction perfectly fits the data, while $R^2 = 0$ indicates that the model performs as well as the naive predictor based on the mean value of the sample set. Equation 2 can yield negative values for R^2 when fitting a non-linear model, in cases when the mean of the data provides a better prediction than the model.

Second, we employ the Root Mean Square Error (RMSE), an accuracy metric measuring how well the proposed model fits the data by averaging the distance between the values predicted by a model and the ones actually observed. Last, we utilize a metric proposed in [4], calculating the percentage of predictions that lie within $\pm x\%$ of the actual value y_i . Let e_i denote the absolute difference between the predicted value p_i and the actual value y_i , i.e. $e_i = |y_i - p_i|$.

$Pred(x)$ is then defined as follows:

$$Pred(x) = \frac{|\{i|e_i/y_i < x/100\}|}{n} \quad (3)$$

6. EXPERIMENTAL RESULTS

We performed the following experiments according to the methodology described above on each project independently. The target field, namely the handling time, is calculated in days. In Table 2, we present a general overview of the system performance on the six projects in our data set. For each project the sample size (S. Size) and several accuracy metrics are given. We see that in Project 2 a high value of R^2 was achieved, and in Projects 3 and 6, our implementation cut the mean square error by half, compared to the naive algorithm. Comparing our results to the work done in [4], our approach achieved better results, in both $Pred(25)$ and $Pred(50)$, in all of the projects in the data set. In the corresponding work less than 30% of the predictions lie within 50% range of the actual effort on average, whereas our results show $Pred(50) = 0.42$, a 40% improvement. We see a similar improvement in $Pred(25)$. It is important to mention that the data sample sets' sizes in this study are significantly larger than the projects used in [4]. These results demonstrate that fields other than the Summary and Description should be considered when predicting how long it will take to fix a defect.

Table 2: Summary of performance results

Project	S. Size	R^2	RMSE	$Pred(25)$	$Pred(50)$
1	2659	0.27	4.815	0.195	0.39
2	4000	0.605	4.045	0.405	0.665
3	1088	0.48	4.44	0.32	0.525
4	4000	0.26	2.895	0.16	0.315
5	3794	0.205	4.835	0.145	0.295
6	493	0.48	3.95	0.18	0.325

In Figure 5 we see the learning curve of the projects in our data set. In this particular graph each experiment was conducted three times, i.e. each point represents three experiments done with a given project and a given sample set size, in addition to the cross-validation done in each experiment. We can see that the results are stable when the data set contains more than 1000

defects. Projects 1, 2 and 3 show high values of R^2 for data sets containing more than 500 defects. The differences in results between the projects should be further investigated.

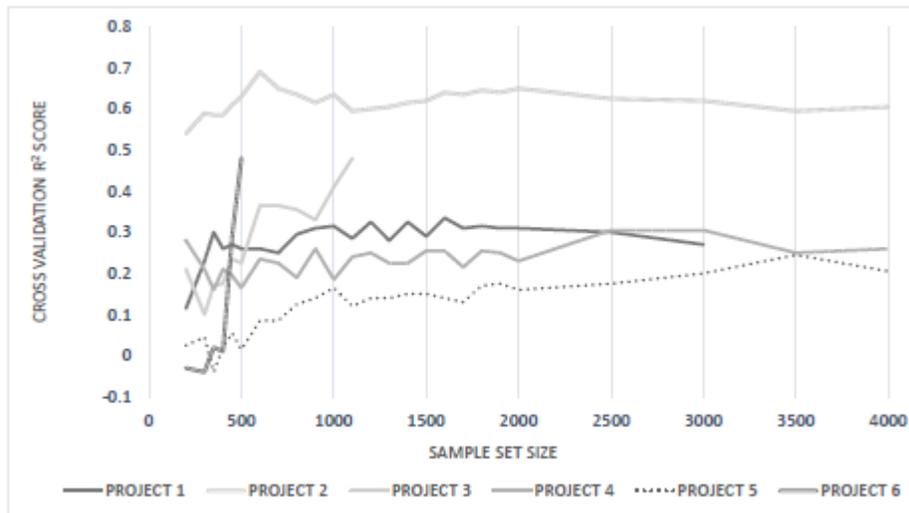
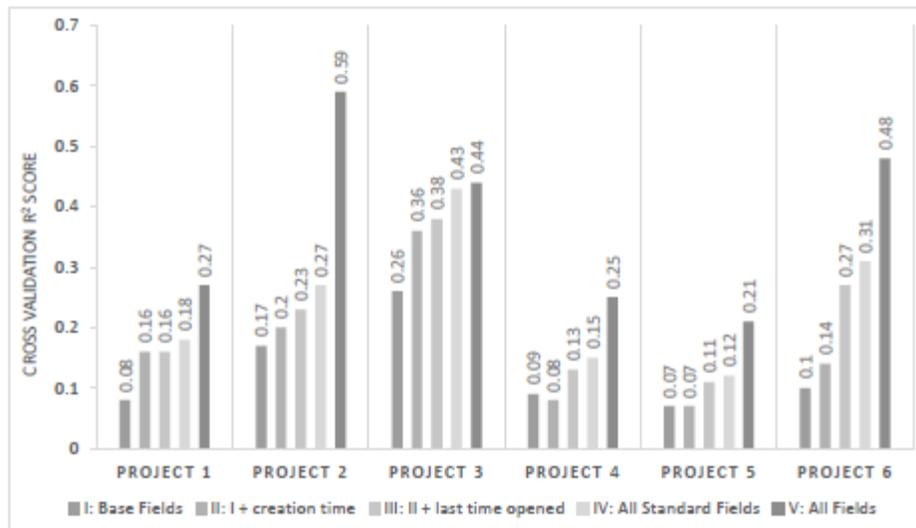


Figure 5: Learning Curve - R^2 as a function of the sample set size

In Figure 6 we compared five sets of data fields for each of the projects used in our experiments. The first set, Base Fields, is a set comprised of the following basic fields: *Detector*, *Assignee*, *Priority*, *Severity* and *Project Name*. The next couple of field-sets were based upon the base set with additional fields specified in the legend of Figure 6; the fourth set included all standard fields, common between all projects; and the last set contained all fields, including user-defined fields, which vary significantly between different projects. As shown in Figure 6, R^2 scores increased in correspondence with increases in the number of fields available for learning, and particularly when the system was allowed to use user-defined fields. It is important to note that any field may or may not be discarded by the algorithm, either in the learning phase itself or during pre-processing.

While analyzing the results and comparing the different projects, we also examined the feature importances (i.e. the weight of each feature computed by the model). This examination showed that user-defined fields played a crucial role in determining the amount of time to fix a defect. In all projects, at least one such field was ranked as one of the three leading features. We also found that the date a defect last entered an Open state is significant for the prediction. This result supports the assumption mentioned briefly in Section , where we explained how the time a defect was opened or detected may be important.

Examining feature importances also showed that fields which are intuitively thought as significant in handling time prediction of defects, are not necessarily so. Fields such as *Severity*, *Priority*, *Assignee* and *Detector* were not found to be noticeably important in most projects. These results support those described in [16].

Figure 6: R² score as a function of the fields used for learning features

To accurately compare our results to the work done in [7], we ran an experiment in which the target field was calculated in hours and the defects were categorized into Fast and Slow defects using the median of fixing time. Similarly, in this experiment we calculated the median for each project and partitioned the defects into two classes, we used the measures described in [7], and used a Random Forest Classifier with parameters close to those used for our regressor (described in Section 1). The results presented in Table 3 are the average over all projects in our dataset, and the average results over all projects presented in the compared work when the initial defect data was used. Their data set contained six open source projects, with a similar number of defects to the current paper. Our results show an improvement over the reported results in the compared study.

Table 3: Performance comparison

	Precision	Recall	AUC
Current	0.742	0.668	0.795
Giger et. al [7]	0.635	0.657	0.702

7. VALIDITY

In this section we discuss the validity of our work by addressing the threats for validity of software engineering research proposed by Yin in [17].

Construct Validity. Our construct validity threats are mainly due to inaccuracies in calculating handling times, based on information stored in defect tracking systems. This is due to the "human factor": developers are expected to update the systems, often manually, to reflect the real work process, and have different expertise levels and variable work availability and productivity.

Internal Validity. Methods such as cross-validation were used to make sure the results are as accurate as possible. In cases where a phenomenon had alternative explanations (e.g. comparison

between results), we tried to incorporate the uncertainties to explain the variability and we are not aware of any other factors that could bias the results.

External Validity. Six different projects from different companies and several industries were examined in this study. Our approach can be applied to almost any data set of defects, with a wide range of diversity. However, a possible threat is that the data sets in this study were all taken from the same implementation of ALM. Further studies on different systems are desirable to verify our findings.

Reliability Validity. The dataset used in the work is commercial so cannot be publicly accessible for the moment and therefore this work cannot be replicated by other researchers using the exact dataset. However, we made a significant effort to provide all relevant implementation details.

8. SUMMARY

In this paper, we presented a novel approach for prediction of software defect handling time by applying data mining techniques on historical defects. We designed and implemented a generic system that extracts defect information, applies preprocessing, and uses a random forest regression algorithm to train a prediction model. We applied our method on six projects of different customers from different industries, with promising results. Our system was designed to handle flaws in the data sets common in real-life scenarios, such as missing and noisy data.

9. FUTURE WORK

We currently do not sufficiently exploit the content of the free-text fields. We would like to use text mining techniques to extract key terms that affect the defect handling time.

We are also considering an expanded approach based on state transition models, in which we calculate the probability of a defect transitioning between any given pair of states during a certain time period. A similar idea was described in [18] but we want to expand this idea by computing a separate model for each defect, based on its fields, rather than assuming that all defects behave identically. This could be used to make a large number of predictions about defect life-cycle, for example, to predict how many defects will be reopened. Combining this with methods used for defect injection rates, such as those surveyed in [19], may provide a more realistic prediction for the situation in which new defects are detected within the project time-frame.

Our algorithm can be easily generalized to other entities representing work to be done, such as product requirements and production incidents, and we would like to evaluate its accuracy in these cases. We plan to also generalize our system to extract data from different defect reporting systems.

To make our data publicly available to be used by related research, we plan to obfuscate our data set by removing any identifiable or private information and publish it.

REFERENCES

- [1] B. Boehm and V. R. Basili, "Software defect reduction top 10 list," *Foundations of empirical software engineering: the legacy of Victor R. Basili*, vol. 426, 2005.
- [2] S. McConnell, *Software estimation: demystifying the black art*. Microsoft press, 2006.
- [3] H. Zeng and D. Rine, "Estimation of software defects fix effort using neural networks," in *Computer Software and Applications Conference, 2004. COMPSAC 2004. Proceedings of the 28th Annual International*, vol. 2, pp. 20{21, IEEE, 2004.
- [4] C. Weiss, R. Premraj, T. Zimmermann, and A. Zeller, "How long will it take to fix this bug?," in *Proceedings of the Fourth International Workshop on Mining Software Repositories*, p. 1, IEEE Computer Society, 2007.
- [5] E. Hatcher, O. Gospodnetic, and M. McCandless, "Lucene in action," 2004.
- [6] R. Hewett and P. Kijsanayothin, "On modeling software defect repair time," *Empirical Software Engineering*, vol. 14, no. 2, pp. 165{186, 2009.
- [7] E. Giger, M. Pinzger, and H. Gall, "Predicting the fix time of bugs," in *Proceedings of the 2nd International Workshop on Recommendation Systems for Software Engineering*, pp. 52{56, ACM, 2010.
- [8] H. Zhang, L. Gong, and S. Versteeg, "Predicting bug-fixing time: an empirical study of commercial software projects," in *Proceedings of the 2013 International Conference on Software Engineering*, pp. 1042{1051, IEEE Press, 2013.
- [9] L. Davies and U. Gather, "The identification of multiple outliers," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782{792, 1993.
- [10] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.
- [11] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5{32, 2001.
- [12] D. S. Hochbaum and D. B. Shmoys, "Using dual approximation algorithms for scheduling problems theoretical and practical results," *Journal of the ACM (JACM)*, vol. 34, no. 1, pp. 144{162, 1987.
- [13] T. Kawaguchi and S. Kyan, "Worst case bound of an lrf schedule for the mean weighted ow-time problem," *SIAM Journal on Computing*, vol. 15, no. 4, pp. 1119{1129, 1986.
- [14] R. L. Graham, "Bounds on multiprocessing timing anomalies," *SIAM journal on Applied Mathematics*, vol. 17, no. 2, pp. 416{429, 1969.
- [15] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137{1145, 1995.
- [16] P. Bhattacharya and I. Neamtiu, "Bug-fix time prediction models: can we do better?," in *Proceedings of the 8th Working Conference on Mining Software Repositories*, pp. 207{210, ACM, 2011.
- [17] R. K. Yin, *Case study research: Design and methods*. Sage publications, 2013.

- [18] J. Wang and H. Zhang, "Predicting defect numbers based on defect state transition models," in Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement, pp. 191-200, ACM, 2012.
- [19] N. E. Fenton and M. Neil, "A critique of software defect prediction models," Software Engineering, IEEE Transactions on, vol. 25, no. 5, pp. 675-689, 1999.

NEED FOR A SOFT DIMENSION

Pradeep Waychal¹ and Luiz Fernando Capretz²

¹College of Engineering, Innovation Center, Pune, India

²Dept. of Electrical and Computer Engineering,
Western University, London, Canada,

ABSTRACT

It is impossible to separate the human factors from software engineering expertise during software development, because software is developed by people and for people. The intangible nature of software has made it a difficult product to successfully create, and an examination of the many reasons for major software system failures show that the reasons for failures eventually come down to human issues. Software developers, immersed as they are in the technological aspect of the product, can quickly learn lessons from technological failures and readily come up with solutions to avoid them in the future, yet they do not learn lessons from human aspects in software engineering. Dealing with human errors is much more difficult for developers and often this aspect is overlooked in the evaluation process as developers move on to issues that they are more comfortable solving. A major reason for this oversight is that software psychology (the softer side) has not developed as extensively

KEYWORDS

Human Factors in Software Engineering, Human Aspects of Engineering, Engineering Accreditation

1. INTRODUCTION

The 2004 version of ACM/IEEE Software Engineering Curriculum mentioned document suggests only 5 hours of studies for group dynamics, whereas the 2013 draft recommends 8 hours. This is clearly not enough for a topic of such crucial importance. Besides, there should be more venues to publish papers with results in this field; workshops and conference sessions can help to increase visibility on this area and to discern the connection between human factors (including the individual, social, cultural, and organizational aspects) and the process of software development. But educators and researchers willing to venture into this area should not face an arduous task if they try to convince their colleagues and software engineering "purists" of the importance of the subject. We need to realize that the human element is pivotal to the engineering of software, and it is worth studying and teaching the soft dimension.

Human-Computer Interaction (HCI), a common course within computer science departments, may be the closest subject and great strides have been made by computing professionals in adopting a human viewpoint to improve user interfaces. Although HCI addresses different topics, as it focuses on people interacting with software, HCI designers have been able to add this new dimension to their design philosophy.

Likewise, software engineers could benefit immensely if even the smallest insights of human factors could be incorporated into their way of thinking. My own experiences as a software engineering educator, manager, and practitioner who continually keep human factors in mind. A course on human factor in software engineering should focus on providing a practical overview of

the software engineering process from a human factor perspective, an alternative within a panorama of technically saturated curricula.

The 2004 version of ACM/IEEE Software Engineering Curriculum mentioned document suggests only 5 hours of studies for group dynamics, whereas the 2013 draft recommends 8 hours. This is clearly not enough for a topic of such crucial importance. Besides, there should be more venues to publish papers with results in this field; workshops and conference sessions can help to increase visibility on this area and to discern the connection between human factors (including the individual, social, cultural, and organizational aspects) and the process of software development. But educators and researchers willing to venture into this area should not face an arduous task if they try to convince their colleagues and software engineering "purists" of the importance of the subject. We need to realize that the human element is pivotal to the engineering of software, and it is worth studying and teaching the soft dimension.

2. ACCREDITATION ATTRIBUTES

Ample work has been reported on the required attributes of software engineering graduates and serious gaps between them and available ones. All the accreditation bodies have cognized that and included a set of attributes that the graduating engineers must have. The specific attributes that are in the limelight are teamwork, critical and creative thinking, ethics, lifelong learning, and communication.

While professional organizations have some qualitative means to evaluate these skills and requisite development programs, the academic environment neither measures nor develops them. This needs to change. We certainly need a systemic approach in both industry and academic environments. That entails a regular cycle of measurement, conceptualization, and execution of development programs. Given the academic load on the students, it is imperative that the development of the attributes is integrated with the core engineering curriculum as recommended by Honor [1]. We will describe our experience in the next few sections.

2.1. Teamwork

Teamwork is involved in virtually every professional activity and therefore should be embedded in every possible academic activity. Colleges and Universities should attempt to bring in as much diversity – in terms of academic performance, social and linguistic background, discipline, gender and culture – as possible in their student teams. They should endeavour to assign different roles to students in different teams so that they can develop varied skills. They should also provide the experience of having interactions across cultures and across disciplines both in physical and virtual modes.

The teamwork skill can be measured using instrument developed by Ohland *et al.* [2] and the one based on "Ten Commandments of Egoless Programming" proposed by Adams [3]. We have been using them and appropriate instructional strategies resulting in students getting immense benefits.

2.2. Creative and Critical Thinking

The engineering field is becoming increasingly complex across all its branches - from traditional civil engineering to modern computer, software, space, mechatronics and genetic engineering. The complexity has increased even more due to a growing interdependence among disciplines and the emergence of a wide range of new technologies. To manage this situation, engineers who are creative and capable of abstract thinking, engineers who can keep pace with new technologies and think laterally when developing new applications are needed. It has been observed that the

recent engineering graduates are lacking in these competencies [4]; and the traditional and still dominant engineering curriculum at most universities, especially in developing countries, makes little provision for developing these competencies [5].

We have been experimenting with Index of Learning Style (ILS)'s [6] sensing intuition preferences to measure critical thinking. We have chosen this instrument over other options like MBTI and TTCT since the latter are costly and have elicited diverse opinions on their suitability. We have also designed a course and found statistically significant changes in critical thinking based on the ILS measure.

2.3. Ethics

This is a complex skill to measure and develop. Its development requires involvement of other stakeholders like parents, society, and industry. Right now, academicians are handling it by introducing a separate traditional or scenario based course. That is not proving to be sufficient as it lacks real life situations with real software engineering stakeholders [7]. We have introduced peer evaluation using constant sum scale for many courses and believe correlations in the self and peer evaluation may provide some idea about ethics of students.

2.4. Life Long Learning

In today's knowledge economy continuous/lifelong learning is assuming greater significance. Every educational institute needs to appreciate that and make provision for developing those skills, for instance to nurture the ability to understand available code written in modern languages in open source software libraries [8].

Colleges require introducing courses or at least some topics in the courses where students have to learn on their own. Performance in them may indicate their skill in the attribute. We have introduced courses on Liberal Learning for all students at the College of Engineering Pune, India. They study various non engineering topics on their own. That provides opportunities to develop lifelong learning skills and a method to evaluate their performance in that important dimension [9].

2.5. Communication

This is critical not only for the engineering career but for all careers. It does not just mean knowledge of English and basic verbal skills but also includes the ability to write emails and proposals, articulate and communicate ideas, create and deliver public presentations, interact across cultures, and listen and analyze talks. This is the easiest of the skills and needs to be developed and measured in every academic course.

For example, engineering software design involves performing tasks in distinct areas, such as system analysis, software design, programming, software testing, and software evolution/maintenance [10]; other software occupations on a design team include the project manager, troubleshooter, helpdesk personnel, database administrator, and so forth. Thus today, specialties within software engineering are as diverse as in any other profession [11]. Therefore, software engineers need to communicate very effectively with users and team members, consequently the people dimension of software engineering is as important as the technical expertise [12].

3. CONCLUSIONS

Software engineering has been doing a marvellous job of helping enterprises of all types and to continue doing so it requires focusing on the soft dimension – people dimension. Its development must start in the undergraduate and graduate courses. The primary areas to work on are: teamwork, creative and critical thinking, ethics, lifelong learning and communication. They still need to be both developed and measured.

REFERENCES

- [1] P. Honor, “Which ABET competencies do engineering graduates find most important in their work?”, *Journal of Engineering Education*, vol. 101, no. 1, January 2012, pp. 95-118.
- [2] M.W. Ohland, R.A. Layton, M.L. Loughry and A.G. Yuhasz, “Effects of behavioral anchors on peer evaluation reliability”, *Journal of Engineering Education*, vol. 94, no. 3, July 2005, pp. 319-332.
- [3] L. Adams, Ten Commandments of Egoless Programming, TechRepublic Blog, 2002, <http://www.techrepublic.com/article/the-ten-commandments-of-egoless-programming/>
- [4] O. Madara, and G.E. Okudan, "Integrating systematic creativity into first-year engineering design curriculum." *International Journal of Engineering Education*, 2006, vol. 22, no. 1, 2006, pp. 109.
- [5] S. Olson, *Educating Engineers: Preparing 21st Century Leaders in the Context of New Modes of Learning: Summary of a Forum*, National Academy of Engineering ISBN 978-0-309-26770-0, 2013, pp. 9.
- [6] R. M. Felder and L.K. Silverman, *Index of Learning Styles (ILS)*, North Carolina State University, 2002, <http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSpage.html>
- [7] B. Berenbach and M. Broy, “Professional and ethical dilemmas in software engineering,” *IEEE Computer*, vol. 42, no. 1, January 2009, pp. 74-80.
- [8] A. Raza, L.F. Capretz and F. Ahmed, “Improvement of Open Source Software Usability: An Empirical Evaluation from Developers Perspective”, *Advances in Software Engineering*, Vol. 2010, pp. 1-12, 2010, DOI: 10.1155/2010/517532.
- [9] P.K. Waychal, “Developing creativity competency of engineers”, *ASEE Annual Conference*, Indianapolis, June 2014, paper ID #8802, pp. 1-15.
- [10] L.F. Capretz, “A Component-Based Software Process”. Wang Y., Patel S. and Johnston, R.H. (editors), in book: *Object-Oriented Information Systems*, Springer, London, U.K., pp. 523-529, 2001, ISBN 1-85233-546-7.
- [11] L.F. Capretz and F. Ahmed, “Making sense of software development and personality types,” *IEEE IT Professional*, vol. 12, no. 1, January 2010, pp. 6-13. DOI: 10.1109/MITP.2010.33.
- [12] F. Ahmed, L.F. Capretz, S. Bouktif and P. Campbell, “Soft Skills and Software Development: A Reflection from the Software Industry”, *International Journal of Information Processing and Management*, vol. 4, no. 3, pp. 171-191, May 2013, DOI: 10.4156/ijipm.vol14.issue3.17.

AUTHORS

Pradeep Waychal chairs an NGO that works at the intersection of human sciences and software engineering, engineering education and innovation management. He has done Ph.D. in developing innovation competencies for Information System Businesses from IIT Bombay, India. He is a senior member of IEEE, a member of ASEE and a life member of CSI – India.

Luiz Fernando Capretz is a professor of software engineering and assistant dean (IT & e-Learning) at Western University in Canada, where he also directed a fully accredited software engineering program. He has vast experience in the engineering of software and is a licensed professional engineer in Ontario. Contact him at lcapretz@uwo.ca or via www.eng.uwo.ca/people/lcapretz/.

INTENTIONAL BLANK

THE ANNUAL REPORT ALGORITHM: RETRIEVAL OF FINANCIAL STATEMENTS AND EXTRACTION OF TEXTUAL INFORMATION

Jörg Hering

Department of Accounting and Auditing, University of Erlangen-Nürnberg,
Lange Gasse 20, Nürnberg D-90403, Germany

ABSTRACT

U.S. corporations are obligated to file financial statements with the U.S. Securities and Exchange Commission (SEC). The SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system containing millions of financial statements is one of the most important sources of corporate information available. The paper illustrates which financial statements are publicly available by analyzing the entire SEC EDGAR database since its implementation in 1993. It shows how to retrieve financial statements in a fast and efficient way from EDGAR. The key contribution however is a platform-independent algorithm for business and research purposes designed to extract textual information embedded in financial statements. The dynamic extraction algorithm capable of identifying structural changes within financial statements is applied to more than 180,000 annual reports on Form 10-K filed with the SEC for descriptive statistics and validation purposes.

KEYWORDS

Textual analysis, Textual sentiment, 10-K parsing rules, Information extraction, EDGAR search engine

1. INTRODUCTION

Information Extraction (IE) can be defined as the process of “finding and extracting useful information in unstructured text” [1]. In contrast to Information Retrieval (IR), a technology that selects a relevant subset of documents from a larger set, IE extracts information from the actual text of documents [2]. Important sources for IE are unstructured natural language documents or structured databases [3] [4]. Since U.S. corporations are obligated by law to file financial statements on a regular basis with the U.S. Securities and Exchange Commission (SEC), the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system containing millions of financial statements is one of the most important sources of corporate information available [5] [1]. Unfortunately, most of the available textual data in the SEC EDGAR database is weakly structured in technical terms [6] [7] [8] especially prior to 2002 when the use of markup

languages was less common [9]. A limited number of tagged items, formatting errors and other inconsistencies lead to difficulties in accurately identifying and parsing common textual subjects across multiple filings [10] [11] [7]. These issues directly affect the ability to automate the extraction of textual information from SEC submissions [10] [12] [13]. Business data providers are offering expensive commercial products (e.g. AcademicEDGAR+, Edgar Pro, Intelligize). As research in the context of textual analysis is growing (e.g. Tetlock 2007 [14]; Loughran and McDonald 2011a [15]; Jegadeesh and Wu 2013 [16]) the question occurs which particular financial statements and disclosures are publicly available for free, how to retrieve these corporate documents and how to decode the embedded textual information in order to be incorporated into investment decisions, trading strategies and research studies in financial economics [5]. Today only a very limited amount of specific literature for extracting textual information from financial statements filed with the SEC and its EDGAR system is available (except Gerdes 2003 [10]; Stümpert et al. 2004 [17]; Grant and Conlon 2006 [1]; Engelberg and Sankaraguruswamy 2007 [18]; Cong, Kogan and Vasarhelyi 2007 [19]; Thai et al. 2008 [20]; Chakraborty and Vasarhelyi 2010 [21]; Hernandez et al. 2010 [22]; Garcia and Norli 2012 [5]; Srivastava 2016 [23]). This paper is based on neither of these because first, non-specialist technology is used to retrieve financial statements in an efficient way and secondly, the algorithm designed to extract textual information is platform-independent. The suggested method can compensate for expensive commercial products and help to replicate empirical research results. The paper shall serve as a technical guide on how to retrieve financial statements filed with the SEC and how to decode the embedded textual information provided by the EDGAR system for business and research purposes.

The remainder of the paper proceeds as follows. Section 2 presents the amount and variety of corporate documents distributed by the SEC's electronic disclosure system. Section 3 demonstrates how to retrieve these documents from the EDGAR database. Section 4 describes the fundamentals of HyperText Markup Language and examines the electronic data provided by the SEC. Section 5 describes the fundamentals of regular expressions and specifies an algorithm to extract textual information embedded in financial statements. Section 6 validates the capabilities of the extraction algorithm. Section 7 presents descriptive statistics of annual reports filed with the EDGAR database. The last section concludes.

2. SEC'S EDGAR DATABASE

Publicly owned companies, their officers and directors as well as major investors are obligated by law (Securities Exchange Act 1934, Section 2) to file various disclosures (forms) with the SEC [10]. The main purpose of making certain types of corporate information publicly available is to improve the efficiency of security markets and to protect capital market participants [5]. "The laws and rules that govern the securities industry in the United States derive from a simple and straightforward concept: all investors, whether large institutions or private individuals, should have access to certain basic facts about an investment prior to buying it, and so long as they hold it. To achieve this, the SEC requires public companies to disclose meaningful financial and other information to the public. This provides a common pool of knowledge for all investors to use to judge for themselves whether to buy, sell, or hold a particular security" [24]. In order to protect investors, to maintain efficient capital markets and to improve access to publicly available corporate disclosures, the SEC developed the EDGAR database [10] and describes it as a system which "performs automated collection, validation, indexing, acceptance, and forwarding of

submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission” [25].

Originally the EDGAR system was developed by the SEC as a pilot system for electronic disclosure in 1983. In order to test and evaluate EDGAR’s performance the SEC requested electronic filings in 1994 after completing the phase-in of a mandated test group in December 1993 (the phase-in began on April 26, 1993) [26] [11] [27]. As of May 6, 1996 the SEC obligated all public domestic U.S. companies (issuers) to file submissions electronically through the EDGAR system [28] [11] [27] [1] except for certain filings made in paper because of a hardship exemption under Regulation S-T [29] [25]. Filing for foreign private issuers (companies organized outside of the U.S.) and foreign governments via EDGAR [26] became mandatory on May 14, 2002 [30]. The Securities Exchange Act of 1934 (Securities Exchange Act 1934, Section 13(a), (b), Section 15(d)) empowers the SEC to require (periodic) reporting of information from publicly held companies [24]. In general, all public domestic companies with assets exceeding \$10 million and at least 500 shareholders become subject to Exchange Act reporting requirements (Securities Exchange Act 1934, Section 12(g)) alongside certain individuals [10]. Among other disclosures, corporations with publicly traded securities are required (Securities Exchange Act 1934, Section 13(a), (b), Section 15(d)) to file annual and quarterly reports (Form 10-K, Form 10-Q) as well as current reports (Form 8-K) on an ongoing basis with the SEC and its EDGAR system [24]. Since by law these public corporate disclosures have to be accurate (Securities Exchange Act 1934, Section 13(i)) and represent a company’s operations, they themselves represent a treasure trove of valuable information for investors and researchers [10] [18].

2.1. Underlying data in SEC’s EDGAR database

In order to understand the amount and variety of corporate information (e.g. financial statements) distributed by the SEC, I retrieve and analyze all form index files since the implementation of the EDGAR system in 1993. The SEC EDGAR form index files list all publicly available disclosures made through the system in a certain quarter and sort the submissions by their particular filing form type. Table 1 reports the total number of submissions that have been made with the EDGAR system for each quarter and year since the introduction of the EDGAR database.

A tremendous amount of publicly available disclosures was filed with the SEC between 1993 and 2016. In total 15,998,058 filings were submitted to the EDGAR system in order to be publicly distributed. On average 31.48 percent (5,035,554) of these filings became available in the first, 25.74 percent (4,117,631) in the second, 20.97 percent (3,355,412) in the third and 21.81 percent (3,489,461) in the last quarter of each year since 1993. Most noticeable is the overall increase in total submissions through the EDGAR system reaching its peak in 2007 with more than 1.1 million disclosures for that particular year. By analyzing the index files more precisely, investors and researchers can gain an insight into the specific type of information the SEC is making publicly available through its EDGAR system [5]. Table 2 describes the most common filing (form) types filed with the EDGAR system.

Table 1. Statistics on EDGAR submissions

Year	Filings (Number)				Filings (Number)	Filings (%)
	Quarter 1	Quarter 2	Quarter 3	Quarter 4		
2016	307,416	239,528	---	---	546,944	3.42
2015	318,519	259,852	206,628	209,216	994,215	6.21
2014	311,679	252,333	212,352	220,328	996,692	6.23
2013	303,568	257,597	213,031	216,266	990,462	6.19
2012	309,453	246,776	203,723	214,985	974,937	6.09
2011	307,644	262,218	207,142	202,628	979,632	6.12
2010	300,538	255,180	203,920	220,070	979,708	6.12
2009	300,080	229,347	200,688	208,396	938,511	5.87
2008	328,709	267,722	220,732	219,669	1,036,832	6.48
2007	339,872	289,082	252,071	256,460	1,137,485	7.11
2006	335,577	278,960	232,131	249,956	1,096,624	6.85
2005	317,761	271,632	242,173	240,725	1,072,291	6.70
2004	312,029	253,021	217,726	241,435	1,024,211	6.40
2003	183,595	167,119	212,258	227,800	790,772	4.94
2002	125,189	108,013	97,533	118,149	448,884	2.81
2001	111,740	90,283	74,313	75,107	351,443	2.20
2000	116,209	81,129	72,571	72,053	341,962	2.14
1999	105,531	78,272	68,631	68,828	321,262	2.01
1998	106,666	73,830	67,234	65,570	313,300	1.96
1997	91,096	65,470	60,142	63,422	280,130	1.75
1996	49,925	47,659	50,641	54,389	202,614	1.27
1995	31,875	26,104	26,699	28,973	113,651	0.71
1994	20,879	16,500	13,066	15,016	65,461	0.41
1993	4	4	7	20	35	0.00
Filings (Number)	5,035,554	4,117,631	3,355,412	3,489,461	15,998,058	100.00
Filings (%)	31.48	25.74	20.97	21.81	100.00	

Notes: The table presents the total number of filings made on EDGAR for each year between 1993 and 2016. Each individual filing in a particular quarter is listed in an associated EDGAR form index file on the SEC server.

Table 2. Statistics on EDGAR form types

Rank	Form/Description	Submission Type	Filings (Number)	Filings (%)
1	Changes in ownership	4	5,850,937	36.57
2	Current report filing	8-K	1,376,248	8.60
3	5% passive ownership triggers amendments	SC 13G/A	587,711	3.67
4	Initial ownership report	3	538,228	3.36
5	Quarterly report	10-Q	522,906	3.27
6	Definitive materials	497	365,987	2.29
7	5% passive ownership triggers	SC 13G	344,030	2.15
8	Current report of foreign issuer	6-K	326,751	2.04
9	Change on a prospectus	424B3	254,046	1.59
10	5% active ownership triggers amendments	SC 13D/A	201,938	1.26
11	Changes in ownership amendments	4/A	197,612	1.24
12	Quarterly holdings, institutional managers	13F-HR	193,463	1.21
13	Annual report on ownership changes	5	186,884	1.17
14	Annual report	10-K	167,599	1.05
15	SEC-originated letters to filers	UPLOAD	159,065	0.99
16	Filer response letters	CORRESP	153,987	0.96
17	Proxy statements	DEF 14A	152,216	0.95
18	Registration management investment companies	485BPOS	151,903	0.95
19	Registration of securities, investment companies	24F-2NT	149,385	0.93
20	Offering of securities without registration	D	147,355	0.92
...
Total			15,998,058	100.00

Notes: The table presents the most frequent form types filed with the EDGAR system between 1993 and 2016. The first column ranks each filing type in descending order of total submissions. The second column gives a short description of each filing form type [5]. The third column lists the form codes used on EDGAR to identify a particular filing type made with the database. The next column contains the number of total submissions of a particular filing form type. The last column shows the amount of total submissions for each filing type in relation to all submissions made with the SEC EDGAR database.

The submission type most often filed with the EDGAR system since its implementation is Form 4. Between 1993 and 2016 5,850,937 filings report purchases or sales of securities by persons who are the beneficial owner of more than ten percent of any class of any equity security, or who are directors or officers of the issuer of the security [5]. The second most frequent submission type filed with the SEC is Form 8-K. 1,376,248 filings of this submission type are listed in the EDGAR index files. The current report filing is required by companies in order to inform shareholders about certain corporate events. These events of material importance for a company include information on significant agreements, impairments, changes in management etc. [5]. Important submission types for investors and researchers such as the annual report on Form 10-K have been submitted 167,599 times. Quarterly reports on Form 10-Q have been filed 522,906 times in total between 1993 and 2016. Another important submission type is Schedule 13G (SC 13G). Investors who are not seeking control over a firm (passive investors) must file this submission type as required by the SEC when crossing the five percent ownership threshold of a company [5]. In total 344,030 filings of this particular submission type alone are reported on EDGAR.

The SEC assigns to each filer a Central Index Key (CIK) which is a unique identifier used on the EDGAR database in order to label and identify each individual filer in the system [10]. Since 1993 in total 580,225 unique CIK numbers were assigned and stored in the SEC's electronic disclosure system. The majority of these CIKs were not assigned to publicly traded companies but to private firms, hedge funds and mutual funds as well as to private individuals who receive a CIK when filing with the SEC [5]. Table 3 reports the number of unique CIKs (unique filers) filing a certain submission type with the SEC and its EDGAR system.

Submission type Form 4 (Form 3) was submitted by 206,652 (187,366) different filers between 1993 and 2016. Annual reports on Form 10-K were submitted to the SEC by 33,968 filers. Quarterly reports on Form 10-Q can be associated with 26,271 unique filers whereas the number of CIKs assigned to current reports on Form 8-K is 38,713. On average each registrant filed 4.9 annual reports on Form 10-K and 19.9 quarterly reports on Form 10 Q with the EDGAR system in addition to 35.6 current reports on Form 8-K since 1993. AFS SenSub Corp. (CIK 1347185), an issuer of asset-backed securities, filed 107 annual reports on Form 10-K (56 on 10-K/A). PowerShares DB Multi-Sector Commodity Trust (CIK 1367306), an investment company offering several investment funds, filed 189 quarterly reports on Form 10-Q (7 on 10-Q/A). Chase Bank USA, National Association (CIK 869090) filed 1,484 Form 8-K statements (12 on 8-K/A). 730 Schedule 13D Forms were filed by Gamco Investors, INC. (CIK 807249), an investment advisory and brokerage service firm, (5,528 on SC 13D/A) whereas FMR LLC (CIK 315066), the financial services conglomerate known as Fidelity Investments, filed 7,726 Schedule 13G Forms (25,447 on SC 13G/A).

Table 3. Statistics on EDGAR filers

Rank	Form/ Description	Submission Type	Unique CIKs	Mean	Med.	Max.
1	Changes in ownership	4	206,652	28.3	7	12,170
2	Initial ownership report	3	187,366	2.9	1	550
3	Offering of securities without registration	D	104,853	1.4	1	375
4	Regulation D exemption filing (paper submission)	REGDEX	87,285	1.5	1	150
5	Changes in ownership amendments	4/A	62,099	3.2	1	338
6	Annual report on ownership changes	5	47,466	3.9	1	473
7	Change on a prospectus	424B3	45,204	5.6	2	9,911
8	5% active ownership triggers	SC 13D	43,381	2.3	1	730
9	5% passive ownership triggers	SC 13G	41,629	8.3	2	7,726
10	Notification of effectiveness for Securities Act registration statement	EFFECT	40,485	2.4	1	86
11	Registration of securities issued in business combination transactions	S-4	40,139	2.0	1	70
12	Current report filing	8-K	38,713	35.6	10	1,484
13	Offering of securities without registration amendments	D/A	35,673	2.8	2	1,601
14	Registration of securities issued in business combination transactions amendments	S-4/A	35,158	2.8	2	63
15	Annual report	10-K	33,968	4.9	3	107
16	5% passive ownership triggers amendments	SC 13G/A	33,339	17.6	4	25,447
17	SEC-originated letters to filers	UPLOAD	31,720	5.0	3	91
18	Filer response letters	CORRESP	30,031	5.1	3	157
19	5% active ownership triggers amendments	SC 13D/A	29,742	6.8	3	5,528
20	Quarterly report	10-Q	26,271	19.9	14	189

Notes: The table presents the most frequent submission types made on EDGAR in descending order of unique SEC registrants filing a particular submission type. The time period is 1993-2016. The fourth column contains the total number of unique filers submitting a particular form type. Columns 5-7 present the means, medians and maxima of particular filing form types submitted by unique SEC filers

3. SEC EDGAR DATA GATHERING

Researchers in the field of finance and accounting often rely on programming languages (Perl, Python, R, SAS, and SPSS) to retrieve financial statements filed with the SEC. The use of a programming language as a tool is problematic for several reasons. First, many people analyzing financial reports are not familiar with these programming languages. For them it is time-consuming to apply a specific and complex coding language to obtain the corporate filings from EDGAR. Secondly, due to downloading only one filing at a time the procedure is very slow especially when obtaining massive data from the database. Thirdly, since incremental changes have to be made to the algorithm to retrieve another filing form type or filings from another company this particular method is very error-prone.

In contrast, widely used internet browsers (e.g. Mozilla-Firefox, Google-Chrome) can be easily equipped with powerful applications (e.g. DownThemAll, GetThemAll) which offer advanced download capabilities. These fully integrated browser extensions are able to identify links contained in a webpage or file and download the desired document parts simultaneously. To feed these applications only a standard MS Excel spreadsheet is necessary. Every filing made through the EDGAR system in a particular quarter between 1993 and 2016 is stored in an associated index file (file extension *.idx) [5]. The EDGAR index files therefore represent a helpful resource in retrieving massive data from the database. They list important information for each filing such as

the name of the filer, the particular central index key, the date and the type of the submission as well as the particular name of the document on the SEC server. In general, four different types of index files are available sorting the filings made on EDGAR by company name, form type, central index key or by submissions containing financial statements formatted in eXtensible Business Reporting Language (XBRL) [31] [32]. When examining the form index files more precisely one can see that the index files do not only contain the name of any filing made on EDGAR but rather the (entire) server path. Table 4 illustrates an excerpt of information stated in the SEC EDGAR form index file from the first quarter of 2016. By opening the index files for example with a simple MS Excel spreadsheet (file extension *.xlsx) a Uniform Resource Locator (URL) can be created for each financial statement which is listed in a particular index file since the name of the filing and its (partial) server path (directory) is stated. To do so the protocol (https://), the hostname (www.sec.gov/) and a link to the archives directory (Archives/) have to be added to the file name from the index file. Table 5 illustrates the URL components of Coca Cola's 2015 annual report on Form 10-K filed with the SEC on February 25, 2016. These URLs which have been composed based on the EDGAR index files can be copied into a plain text file (file extension *.txt). By opening it with the browser extensive data (financial statements) can be retrieved from the SEC and its EDGAR system in a fast and efficient way using a browser extension (however, the composed URLs can also be implemented in any other data gathering method).

This method offers various significant advantages. First, for many people composing URLs with commonly used and easy accessible computer software like MS Excel is simpler and faster than relying on complex coding languages to identify and retrieve the documents in question. Secondly, since multiple documents can be retrieved at the same time using browser extensions, the described method is again a lot faster especially when obtaining massive data from EDGAR. Thirdly, by sorting or filtering the different index files in MS Excel the proposed method can easily be adjusted to retrieve another filing form type or data from another company. The result of this procedure is validated through obtaining exactly the same financial statements investors and researchers would retrieve using a complex, slow and error-prone alternative.

4. HYPERTEXT MARKUP LANGUAGE IN SEC FILINGS

Because financial statements filed with the SEC are formatted in HyperText Markup Language (HTML) the fundamentals of HTML are illustrated first, followed by an examination of the data formatted in HTML provided by the SEC and its EDGAR system.

4.1. Fundamentals of HyperText Markup Language

HyperText Markup Language (HTML) is a universally understood digital language which is used to publish and distribute information globally. HTML is the publishing language of the World Wide Web [33]. HTML is used to create HyperText documents that are portable from one platform to another [34] due to their generic semantics as a Standard Generalized Markup Language (SGML) application [33]. HTML enables authors to publish documents online, assign a specific look or layout to document content (tagging) [35] [21] or to retrieve information online via HyperText links [33]. The World Wide Web Consortium (W3C) is maintaining and specifying the vocabulary (applicable markups) and grammar (logical structure) of HTML documents [35].

A valid HTML document is composed of three different parts [33]. First, it declares which version of HTML is used in the document through the document type declaration (<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">). The document type declaration names the document type definition (DTD) specifying which elements and attributes can be implemented into a document formatted in HTML [33]. HTML 4.01 specifies three different DTDs: HTML 4.01 Strict DTD; HTML 4.01 Transitional DTD and HTML 4.01 Frameset DTD [33]. The W3C recommends to use HTML 4.01 Strict DTD which excludes presentation attributes since these elements are supposed to be replaced by style sheets [36]. The second part of a HTML document is the document head (<HEAD>). This section contains information about the current document such as the title and relevant keywords for search engines. In general, the elements appearing in the head section are not presented by a document formatted in HTML [33]. The third and most important part of a HTML document is the body (<BODY>). This section contains the actual content of the document such as text paragraphs, images, graphics, tables, links, etc. [33]. The content in the document body can be structured in many different ways using various HTML elements (tags) to accomplish a certain look or layout to present the embedded information.

4.2. SEC EDGAR HTML Data

“Official” financial statements filed with the SEC have to be formatted either in American Standard Code for Information Interchange (ASCII) or in HyperText Markup Language (HTML 3.2/4.0). Financial statements formatted in Portable Document Format (PDF) or XBRL are considered “unofficial” documents (submissions formatted in PDF and XBRL may qualify as official documents as well when specific criteria are met) [34]. Due to a limited support of HTML in order to reduce the number of inconsistencies caused by HTML 4.0 implementation variances [37], the EDGAR system only accepts a subset of HTML 3.2 semantics (tags) and several HTML 4.0 attributes [34] therefore enforcing several restrictions (no active content, no external references etc.) of HTML formatting in financial statement submissions [34]. The “Complete Submission Text File” (file extension *.txt) provided by the EDGAR system represents an aggregation of all information in a particular financial statement filed with the SEC. The text version of the filings on the SEC server contains the 10-K document formatted in HTML, XBRL, exhibits and ASCII-encoded graphics (“binary-to-text” encoding or “uuencoding” converts binary data files to plain ASCII-printable characters to facilitate transfer across various hardware platforms) [38] [39]. Besides the “Complete Submission Text File” several submission parts (documents) are also provided in HTML (file extension *.htm) such as the core 10-K document and the exhibits which have been submitted [38]. For example, Coca Cola’s 10-K filing on February 25, 2016 lists the core 10-K filing in HTML format, ten exhibits, eight graphic files (file extension *.jpg), six XBRL files and a single “Complete Submission Text File” containing all of these documents [40].

5. TEXTUAL INFORMATION IN FINANCIAL STATEMENTS

This section describes how regular expressions are used to extract textual information from financial statements filed with the SEC. First, I illustrate the fundamentals of regular expressions. Then I discuss the algorithm to extract textual information from financial statements using only regular expressions before presenting the actual text embedded in financial statements as a result of the designed algorithm. Due to their high relevance for investors and researchers an actual annual report on Form 10-K from the Coca Cola Company serves as basis for the illustration

5.1. Fundamentals of Regular Expressions

Regular expressions or regular sets were first used as an algebra by mathematicians to describe models developed by physiologists of how the nervous system would work at the neuron level. The first published computational use of regular expressions was in 1968 by Ken Thompson [41] who describes regular expressions as “a method for locating specific character strings embedded in character text” [42]. They are implemented not only in modern programming languages, but also in application programs that can be used for text analysis without special programming skills (e.g. RapidMiner).

Regular expressions (“RegEx”; “RegExp”; “RegExes”) with a general pattern notation (pattern language) allow to process all kinds of text and data in a flexible and efficient way [41] [13]. In particular RegExes can be used to modify textual elements or to identify and extract certain information from different documents [43]. The two types (full) regular expressions are composed of special characters (metacharacters) and normal (literal) text characters acting as the grammar and the words of the regular expression language [41] [43]. For example, RegEx: “[0-9]” identifies all digits, RegEx: “[a-zA-Z]” isolates all upper and lower-case letters (character classes) and RegEx: “.” matches all of these elements (metacharacter) embedded in an underlying text document [41] [43]. Another metacharacter and counting element (quantifier) within the regular expression language is a star or an asterisk (*) which quantifies the immediately preceding item within the defined expression (match any number of the preceding element including none) [41] [43]. Counting elements or quantifiers are used to specify the search pattern of regular expressions in more detail. “Greedy” quantifiers like “*” match as much as possible whereas “lazy” quantifiers such as “*?” match as little as possible to satisfy the search pattern of a composed regular expression [41] [43].

In addition, regular expressions can be modified in the way they are interpreted and applied using different regular expression modes (modifiers). These modifiers allow to change the search pattern of a particular regular expression (matching mode) in modern programming languages or in application programs. Regular expressions equipped with “case-insensitive match mode” ((?i)) ignore the letter case of the input (textual elements) during the matching process allowing the search pattern to match both upper and lower case letters [41] [43]. Since modern applications work with multiple (coding) lines regular expressions need to be modified in order to match a string across different lines. “Dot-matches-all match mode” also known as “single-line mode” ((?s)) modifies the search pattern of a regular expression in a way that it matches a character string across multiple lines [41] [43]. By designing regular expressions and implementing them into modern computer software the results of various search patterns (textual information) can be highlighted and changed or even removed from the underlying text at all [41] [43].

5.2. Extraction of Textual Information

Researchers in the field of finance and accounting (as well as business data providers) use the “Complete Submission Text Files” (file extension *.txt) provided by the SEC and its EDGAR system to extract textual information from financial statements. In order to delete all non-textual elements (HTML tags and their corresponding attributes) most often special text-processing programs and their predefined applications (HTML Parser) are used. This again is problematic for several reasons. First, using predefined text-processing operators to delete non-textual elements makes one platform-dependent since a specific HTML Parser can not be (easily) implemented into any other text-processing program in use. Secondly, since the extraction algorithm of the HTML-Parser is complex or not presented at all its extraction results can hardly be validated. Thirdly, because of these drawbacks empirical research results are challenging to replicate for a particular or any other data sample. Regular expressions can in fact overcome these problems in extracting textual information

embedded in financial statements filed with the SEC. They offer platform-independent (research) results which can be validated and replicated for any data sample at any given time.

The proposed extraction algorithm (“Annual Report Algorithm”) first decomposes the “Complete Submission Text File” (file extension *.txt) into its components (RegEx 1). In the end, the entire algorithm is validated through obtaining exactly one core (Form 10-K) document and the number of exhibits which have been embedded in the “Complete Submission Text File” for every financial statement in the data sample. Next, the “Annual Report Algorithm” identifies all other file types contained in the submission since these additional documents are not either a core document or an exhibit within the text version of the filing (RegEx 2). Table 4 illustrates the regular expressions needed to decompose the “Complete Submission Text File” of a financial statement filed with the SEC and to identify the embedded document (file) types.

Table 4. Regular expressions contained in the “Annual Report Algorithm”

ID	Description	Regular Expression
1	Decomposition of “Complete Submission Text File”	(?s)<DOCUMENT>.*?</DOCUMENT>
2	Identification of document (file) types	<TYPE>.*

Notes: The table presents the regular expressions contained in the “Annual Report Algorithm” for extracting documents and identifying document (file) types.

In addition to the filing components described earlier (10-K section, exhibits, XBRL, graphics), several other document (file) types might be embedded in financial statements such as MS Excel files (file extension *.xlsx), ZIP files (file extension *.zip) and encoded PDF files (file extension *.pdf). By applying additional rules in the “Annual Report Algorithm” (RegExes 3-22) these documents are deleted to be able to extract textual information only from the core document and the various exhibits contained in the “Complete Submission Text File”. The additional SEC-header is not supposed to be removed separately since it has already been deleted by the algorithm. Table 5 illustrates the regular expressions applied to delete document (file) types other than the core document and the corresponding exhibits.

Table 5. Regular expressions contained in the “Annual Report Algorithm”

ID	Description	Regular Expression
3	Removal of graphic files	(?s)<TYPE>GRAPHIC.*?</TEXT>
4	Removal of MS Excel files	(?s)<TYPE>EXCEL.*?</TEXT>
5	Removal of PDF files	(?s)<TYPE>PDF.*?</TEXT>
6	Removal of ZIP files	(?s)<TYPE>ZIP.*?</TEXT>
7	Removal of cover letter	(?s)<TYPE>COVER.*?</TEXT>
8	Removal of correspondence	(?s)<TYPE>CORRESP.*?</TEXT>
9	Removal of XBRL instance document	(?s)<TYPE>EX-10[01].INS.*?</TEXT>
10	Removal of XBRL instance document	(?s)<TYPE>EX-99.SDR [KL].INS.*?</TEXT>
11	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-10[01].SCH.*?</TEXT>
12	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-99.SDR [KL].SCH.*?</TEXT>
13	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-10[01].CAL.*?</TEXT>
14	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-99.SDR [KL].CAL.*?</TEXT>
15	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-10[01].DEF.*?</TEXT>
16	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-99.SDR [KL].LAB.*?</TEXT>
17	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-10[01].LAB.*?</TEXT>
18	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-99.SDR [KL].LAB.*?</TEXT>
19	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-10[01].PRE.*?</TEXT>
20	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-99.SDR [KL].PRE.*?</TEXT>
21	Removal of XBRL taxonomy extension	(?s)<TYPE>EX-10[01].REF.*?</TEXT>
22	Removal of XBRL documents	(?s)<TYPE>XML.*?</TEXT>

Notes: The table presents the regular expressions contained in the “Annual Report Algorithm” for deleting nonrelevant document (file) types.

Next, the “Annual Report Algorithm” deletes all metadata included in the core document and the exhibits (RegExes 23-27). Table 6 illustrates the regular expressions for deleting metadata in SEC EDGAR documents.

Table 6. Regular expressions contained in the “Annual Report Algorithm”

ID	Description	Regular Expression
23	Removal of document type information	<TYPE>.*
24	Removal of sequence information	<SEQUENCE>.*
25	Removal of filename	<FILENAME>.*
26	Removal of description	<DESCRIPTION>.*
27	Removal of head section (including document title)	(?s)<HEAD>.*?</HEAD>

Notes: The table presents the regular expressions contained in the “Annual Report Algorithm” for deleting nonrelevant document metadata.

Before deleting all HTML elements and their corresponding attributes (RegEx 29) the algorithm deletes tables since they contain non-textual (quantitative) information (RegEx 28). Table 7 illustrates the set of regular expressions applied to delete tables and HTML elements embedded in financial statements filed with the SEC.

Table 7. Regular expressions contained in the “Annual Report Algorithm”

ID	Description	Regular Expression
28	Removal of table content	(?s)(?i)<Table.*?</Table>
29	Removal of HTML tags and attributes	(?s)<[^>]*>

Notes: The table presents the regular expressions contained in the “Annual Report Algorithm” for deleting tables and HTML elements

After extracting the core document and the exhibits as well as deleting all HTML elements, the “Annual Report Algorithm” adjusts the content embedded in the body section of each HTMLformatted document in order to extract textual elements from financial statements on the EDGAR database. According to the SEC filer manual the EDGAR system suspends financial statements which contain extended ASCII characters. However, it supports submissions with extended character references. By using ISO-8859-1/Latin-1 decimal character references or entity-names (either technique is allowed within SEC submissions) extended ASCII characters can be embedded in financial statement submissions. These extended character sets within HTML documents included in the “Complete Submission Text File” need to be decoded to be able to extract human-readable textual information from financial statements [34]. The “Annual Report Algorithm” finally decodes all extended character sets (RegExes 30-680) most likely embedded in financial statements filed with the SEC and its EDGAR system formatted in HTML 4.01 (ASCII, ANSI/Windows-1252, ISO-8859-1/Latin-1, mathematical, Greek, symbolic and special characters).

5.3. Extraction Results

By applying the “Annual Report Algorithm” investors and researchers are able to extract textual information from financial statements filed with the SEC for thousands of companies in a fully automated process. Based on the “Complete Submission Text File” provided by the EDGAR system the algorithm extracts the core (Form 10-K) document and the exhibits which have been embedded in the text version of a company’s financial statement. For example for Coca Cola’s 2015 annual report on Form 10-K filed on February 25, 2016 via EDGAR the algorithm extracts one core document in

addition to ten different exhibits. Figure 1 illustrates partial extraction results for the 10-K section of the annual report as well as for two exhibits.

<p>UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C. 20549 FORM 10-K For the fiscal year ended December 31, 2015 OR For the transition period from to Commission File No. 001-02217 (Exact name of Registrant as specified in its charter) Registrant's telephone number, including area code: (404) 676-2121 Securities registered pursuant to Section 12(b) of the Act: Securities registered pursuant to Section 12(g) of the Act: None...</p>
<p>Exhibit 23.1 CONSENT OF INDEPENDENT REGISTERED PUBLIC ACCOUNTING FIRM We consent to the incorporation by reference in the registration statements and related prospectuses of The Coca-Cola Company listed below of our reports dated February 25, 2016, with respect to the consolidated financial statements of The Coca-Cola Company and subsidiaries, and the effectiveness of internal control over financial reporting of The Coca-Cola Company and subsidiaries, included in this Annual Report (Form 10-K) for the year ended December 31, 2015. /s/ ERNST & YOUNG LLP Atlanta, Georgia February 25, 2016...</p>
<p>EXHIBIT 31.1 CERTIFICATIONS I, Muhtar Kent, Chairman of the Board of Directors and Chief Executive Officer of The Coca-Cola Company, certify that: 1. I have reviewed this annual report on Form 10-K of The Coca-Cola Company; 2. Based on my knowledge, this report does not contain any untrue statement of a material fact or omit to state a material fact necessary to make the statements made, in light of the circumstances under which such statements were made, not misleading with respect to the period covered by this report...</p>

Figure 1. Examples of the extraction result of the “Annual Report Algorithm“

Notes: The figure presents extraction results from Coca Cola’s 2015 annual report on Form 10-K filed with the SEC. The first part of the figure displays the actual 10-K section embedded in text version of the submission. The second part shows the statement of the auditing firm. The certification of the annual report by the CEO is presented in the last part of the figure.

Besides from textual content of entire documents (10-K section and exhibits) contained in the “Complete Submission Text File” investors and researchers might be interested in extracting textual information from particular sections (Items) within the core 10-K section of an annual report (like Item 1A - Risk Factors; Item 3 - Legal Proceedings; Item 7 - Management’s Discussion and Analysis of Financial Condition and Results of Operations etc.). In order to extract textual information from particular 10-K items the “Annual Report Algorithm” is modified to the “Items Algorithm”. Excluding all exhibits, the modified “Items Algorithm” isolates only the 10 K section within the SEC submission. After deleting nonrelevant information and decoding reserved characters within the document investors and researchers can extract textual information from specific 10-K items. Table 8 specifies the modified “Items Algorithm” applied to extract textual information from particular items of the annual report on Form 10-K filed with the SEC.

Using only regular expressions to extract textual information from financial statements investors and researchers can implement the designed extraction algorithms in any modern application and computer program available today. By applying either the “Annual Report Algorithm” or the “Items Algorithm” entire documents (10-K section and exhibits) or particular items from the core 10-K section can be extracted from the annual SEC submissions in order to be analyzed. More importantly, while compensating for expensive commercial products the algorithms and their extraction results can be validated and replicated for any data sample at any given time. Figure 2 finally illustrates several extraction results of the “Items Algorithm” from the annual report on Form 10 K highly relevant to investors and researchers alike.

Table 8. Regular expressions contained in the “Items Algorithm”

ID	Description	Regular Expression
1.1	Extraction of 10-K section	(?s)-TYPE>10-K.*?</TEXT>
2.1	Removal of document metadata	RegExes 23-28
3.1	Removal of table content	(?s)(?)<Table.*?</Table>
4.1	Decoding of reserved characters	See RegExes 30-680
5.1	Identification and renaming of item headings (“>Item”)	(?s)(?i)(?m)> +Item>Item^Item
6.1	Removal of multiple empty spaces	(?s) +
7.1	Extraction of Item 1. - Business	(?s)(?)°Item 1[^AB012345].*?°Item
7.2	Extraction of Item 1A. - Risk Factors	(?s)(?)°Item 1A.*?°Item
7.3	Extraction of Item 1B. - Unresolved Staff Comments	(?s)(?)°Item 1B.*?°Item
7.4	Extraction of Item 2. - Properties	(?s)(?)°Item 2.*?°Item
7.5	Extraction of Item 3. - Legal Proceedings	(?s)(?)°Item 3.*?°Item
7.6	Extraction of Item 4. - Mine Safety Disclosures	(?s)(?)°Item 4.*?°Item
7.7	Extraction of Item X. - Executive Officers of the Company	(?s)(?)°Item X.*?°Item
7.8	Extraction of Item 5. - Market for Registrant’s Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities	(?s)(?)°Item 5.*?°Item
7.9	Extraction of Item 6. - Selected Financial Data	(?s)(?)°Item 6.*?°Item
7.10	Extraction of Item 7. - Management’s Discussion and Analysis of Financial Condition and Results of Operations	(?s)(?)°Item 7[^A].*?°Item
7.11	Extraction of Item 7A. - Quantitative and Qualitative Disclosures About Market Risk	(?s)(?)°Item 7A.*?°Item
7.12	Extraction of Item 8. - Financial Statements and Supplementary Data	(?s)(?)°Item 8.*?°Item
7.13	Extraction of Item 9. - Changes in and Disagreements with Accountants on Accounting and Financial Disclosure	(?s)(?)°Item 9[^AB].*?°Item
7.14	Extraction of Item 9A. - Controls and Procedures	(?s)(?)°Item 9A.*?°Item
7.15	Extraction of Item 9B. - Other Information	(?s)(?)°Item 9B.*?°Item
7.16	Extraction of Item 10. - Directors, Executive Officers and Corporate Governance	(?s)(?)°Item 10.*?°Item
7.17	Extraction of Item 11. - Executive Compensation	(?s)(?)°Item 11.*?°Item
7.18	Extraction of Item 12. - Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters	(?s)(?)°Item 12.*?°Item
7.19	Extraction of Item 13. - Certain Relationships and Related Transactions, and Director Independence	(?s)(?)°Item 13.*?°Item
7.20	Extraction of Item 14. - Principal Accounting Fees and Services	(?s)(?)°Item 14.*?(°Item</TEXT>)
7.21	Extraction of Item 15. - Exhibits, Financial Statement Schedules	(?s)(?)°Item 15[^°]*?</TEXT>
8.1	Removal of HTML tags and attributes	(?s)-[<?>]*>

Notes: The table presents the regular expressions contained in the modified “Items Algorithm” for extracting particular items from the annual report on Form 10-K. RegExes 1.1-6.1 modify the text version of a financial statement to be able to extract (clear) textual information from particular items. RegExes 7.1-7.21 represent the actual regular expressions designed to extract particular sections from the text version of the annual report.

<p>°Item 1A. RISK FACTORS In addition to the other information set forth in this report, you should carefully consider the following factors, which could materially affect our business, financial condition or results of operations in future periods. The risks described below are not the only risks facing our Company. Additional risks not currently known to us or that we currently deem to be immaterial also may materially adversely affect our business, financial condition or results of operations in future periods...</p>
<p>°Item 3. LEGAL PROCEEDINGS The Company is involved in various legal proceedings, including the proceedings specifically discussed below. Management believes that the total liabilities to the Company that may arise as a result of currently pending legal proceedings will not have a material adverse effect on the Company taken as a whole. Aqua-Chem Litigation On December 20, 2002, the Company filed a lawsuit (The Coca-Cola Company v. Aqua-Chem, Inc., Civil Action No. 2002CV631-50) in the Superior Court of Fulton County, Georgia...</p>
<p>°Item 7. MANAGEMENT’S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS Overview The following Management’s Discussion and Analysis of Financial Condition and Results of Operations (“MD&A”) is intended to help the reader understand The Coca-Cola Company, our operations and our present business environment. MD&A is provided as a supplement to - and should be read in conjunction with - our consolidated financial statements and the accompanying notes thereto contained in "Item 8. Financial Statements and Supplementary Data" of this report. This overview summarizes the MD&A, which includes the following sections...</p>

Figure 2. Examples of the extraction result of the “Items Algorithm”

Notes: The figure presents extraction results from Coca Cola’s 2015 annual report on Form 10-K filed with the SEC. The first part of the figure displays Item 1A (Risk Factors) embedded in the overall 10-K section.

The last two parts of the figure show Item 3 (Legal Proceedings) and Item 7 (Management’s Discussion and Analysis of Financial Condition and Results of Operations) contained in the 10-K section of the “Complete Submission Text File”

6. VALIDATION OF EXTRACTION ALGORITHMS

In order to validate the proposed extraction algorithms and to test their capabilities, I retrieve all Form 10-K filings listed in the SEC EDGAR form index files. Using the data gathering method as described in Section 3 in total 188,875 annual reports (167,599 on Form 10-K and 21,276 on Form 10-K405) filed between 1993 and 2016 are retrieved from the EDGAR database (SEC EDGAR Form 10-K types as used in Loughran and McDonald 2011a). The “Annual Report Algorithm” is applied to all submissions to derive different word counts for each filing made with the SEC. In addition to the overall word count of an annual report, for each core document (10-K section) and the exhibits embedded in a “Complete Submission Text File” an individual word count is retrieved in order to be compared (XBRL files declared as exhibits are deleted). Figure 3 illustrates how word counts for each filing and its components are obtained from the “Complete Submission Text File” for the document validation process of the “Annual Report Algorithm”.

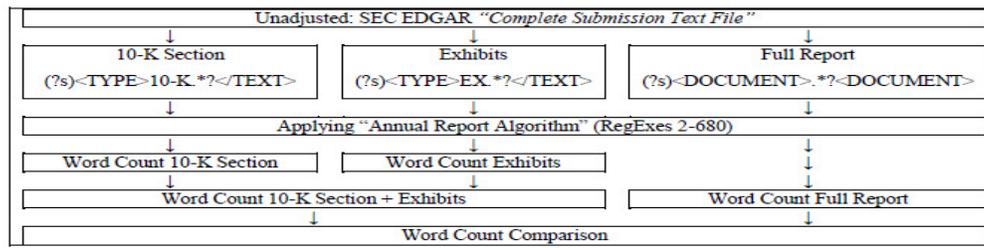


Figure 3. Document validation process of the “Annual Report Algorithm”.

Notes: The figure presents the document validation process of the “Annual Report Algorithm”. The “Complete Submission Text File” of each financial statement as provided on the SEC server is used to extract all relevant components (documents). The “Annual Report Algorithm” is applied to each filing in order to retrieve word counts for all relevant documents embedded in the submission. The word count of all relevant documents is compared with the overall length of the submission. A mismatch between the word counts would indicate that the entire report contains nonrelevant document (file) types after applying the “Annual Report Algorithm”.

This word count comparison between the overall report on full length and its different components cannot be a validation of the “Annual Report Algorithm” since the same algorithm is simply applied to different sets of textual information (10-K section, exhibits, full report). However, if the entire report would still contain document (file) types or elements which are not a part of the core 10-K section or a corresponding exhibit the word count of a certain financial statement would be artificially increased (Word Count Full Report). In fact, the ability to validate the entire extraction procedure by applying an alternative to the “Annual Report Algorithm” (e.g. HTML-Parser) is limited since to a certain extent the same regular expressions have to be used to create the input for both extraction methods in the first place (extracting core 10-K document and exhibits, deleting nonrelevant document (file) types etc.). Due to this disability in validating the entire extraction process from the beginning by applying an HTML-Parser one has to validate the input the proposed algorithm is creating and its extraction results separately, therefore validating the entire information extraction process. The validation of the textual input created by the “Annual Report Algorithm” is represented by the extraction algorithm itself since it uses only regular expressions combined with the electronic filing requirements introduced by the SEC (precisely not the SEC but Attain, LLC). According to the SEC, all documents embedded in a “Complete Submission Text File” must be equipped with a

“<TYPE>” tag representing the conformed document type of that particular submission part within the text version of the filing (<TYPE>10-K, <TYPE>10-Q, <TYPE>8-K, <TYPE>EX-1, <TYPE>EX-2 etc.) [45]. The “Annual Report Algorithm” (RegExes 1-29) uses these requirements in order to extract the core document and the corresponding exhibits from annual reports while deleting all documents associated with XBRL and other document (file) types. The search patterns of the “Annual Report Algorithm” which have been designed accordingly to the filing requirements of the SEC can be validated due to the general pattern notation of the regular expression language.

An output comparison between the “Annual Report Algorithm” and a common HTML-Parser shall serve as an additional validation for the remaining extraction procedure. Therefore, I modify the “Complete Submission Text Files” as provided by the SEC (unadjusted filings) and apply the first part of the “Annual Report Algorithm” (RegExes 1-29) in order to make the text version of the financial statements readable for the predefined HTML-Parser (adjusted filings). Since this part of the overall validation process focuses on how well the “Annual Report Algorithm” is capable of decoding escape sequences embedded in a “Complete Submission Text File” the aggregated text length of both procedures are compared rather than the word counts due to decimal character encodings (a simple word count comparison would not fully capture the disability of the “Annual Report Algorithm” in decoding these character references in relation to the HTML-Parser). Figure 4 illustrates the output validation process of the “Annual Report Algorithm”.

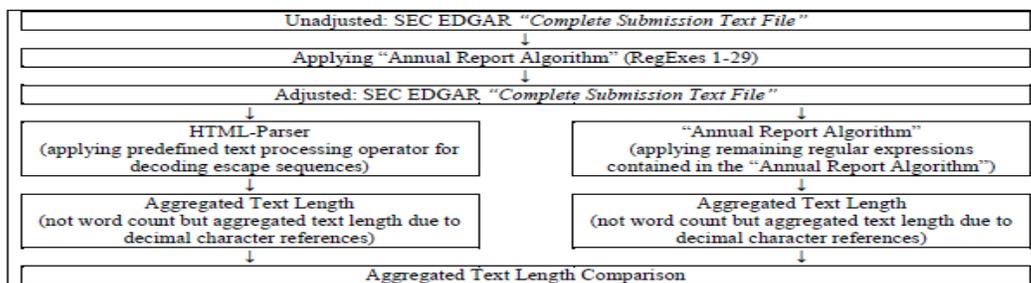


Figure 4. Output validation process of the „Annual Report Algorithm”

Notes: The figure presents the output validation process of the “Annual Report Algorithm”. The “Complete Submission Text File” of each financial statement as provided on the SEC server is adjusted in order to compare the output of the algorithm with the output a common HTMLParser would produce. RegExes 1-29 modify the unadjusted document as provided on the EDGAR database before applying a predefined text processing operator (HTML-Parser). The aggregated text length for all filings of both procedures is compared in order to validate the capability of the „Annual Report Algorithm” in decoding escape sequences. The aggregated text length includes each individual element in an underlying text document (text, digits, spaces, special characters etc.).

In contrast to the “Annual Report Algorithm” the modified “Items Algorithm” is validated by its ability to distribute the extracted information to the individual items an annual report filed with the SEC is composed of. In order to test and validate the capabilities of the “Items Algorithm” I again use the “Complete Submission Text Files” as provided by the SEC and extract only the 10-K section of each filing. For each submission, I retrieve separate word counts for the 10-K section and for all individual items extracted by the “Items Algorithm”. Despite textual information embedded in the 10-K section not contained in a particular item (introduction) a word count comparison between the overall 10-K section and all items represents an attempt to validate the capabilities of the “Items Algorithm” in extracting certain sections from the core document of an annual report filed with the SEC and its EDGAR system. Figure 5 illustrates the content validation process of the “Items Algorithm”.

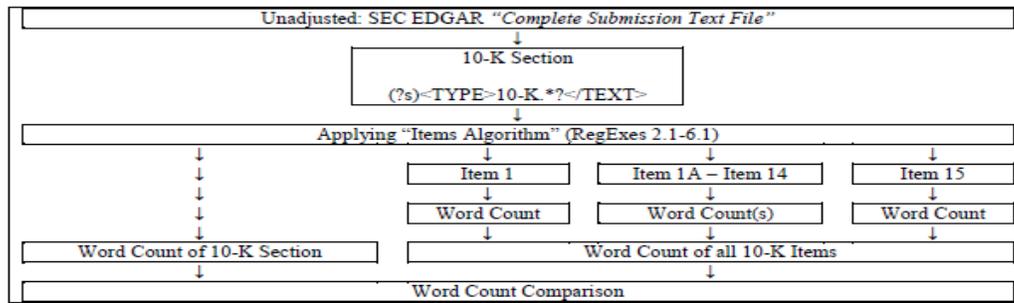


Figure 5. Content validation process of the “Items Algorithm”

Notes: The figure presents the content validation process of the “Items Algorithm”. First, the entire 10-K section of each filing from the “Complete Submission Text File” as provided on the SEC server is extracted. Word counts for the entire 10-K section as well as for all individual items are retrieved by applying the “Items Algorithm” in order to be compared. Due to structural changes of the annual report on Form 10-K over time (different number of items) the relation of text length between the overall 10-K section and all individual items shall represent the ability of the algorithm to extract particular items from the 10-K section.

Table 9 presents the validation results for the “Items Algorithm”.

Year	Filings		“Items Algorithm”								
			Word Count Comparison		Precision, Recall, and F-measure						
	Number	%	Σ of Items (%)	Rest/Error (%)	Filings	Items					
Tested					Exists	Extracted	Correct	Precision	Recall	F-measure	
2016	2,886	44.63	97.72	2.28	10	195	191	185	96.86	94.87	95.85
2015	3,714	46.51	97.60	2.40	10	200	200	195	97.50	97.50	97.50
2014	4,004	49.53	97.52	2.48	10	198	197	193	97.97	97.47	97.72
2013	3,962	48.88	97.53	2.47	10	192	188	188	100.00	97.92	98.95
2012	3,938	46.92	97.39	2.61	10	198	192	183	95.31	92.42	93.85
2011	4,104	46.43	97.43	2.57	10	193	191	189	98.95	97.93	98.44
2010	2,805	30.61	97.10	2.90	10	197	168	155	92.26	78.68	84.93
2009	2,719	27.63	97.15	2.85	10	196	181	164	90.61	83.67	87.00
2008	2,077	23.75	97.28	2.72	10	196	184	170	92.39	86.73	89.47
2007	2,065	24.08	97.38	2.62	10	195	184	173	94.02	88.72	91.29
2006	2,662	30.07	97.49	2.51	10	198	184	165	89.67	83.33	86.39
2005	3,122	34.62	97.70	2.30	10	181	175	163	93.14	90.06	91.57
2004	3,496	40.81	97.60	2.40	10	173	170	163	95.88	94.22	95.04
2003	3,903	46.09	97.33	2.67	10	161	161	154	95.65	95.65	95.65
2002	4,961	55.57	97.65	2.35	10	150	150	150	100.00	100.00	100.00
2001	5,799	62.71	97.61	2.39	10	146	144	135	93.75	92.47	93.10
2000	6,268	63.51	97.55	2.45	10	150	149	138	92.62	92.00	92.31
1999	6,302	62.26	97.55	2.45	10	146	145	143	98.62	97.95	98.28
1998	6,492	63.11	97.56	2.44	10	140	140	128	91.43	91.43	91.43
1997	6,397	64.62	97.43	2.57	10	132	129	125	96.90	94.70	95.79
1996	3,918	62.61	97.27	2.73	10	136	121	112	92.56	82.35	87.16
1995	1,907	58.93	97.07	2.93	10	135	135	127	94.07	94.07	94.07
1994	1,039	54.03	97.23	2.77	10	140	139	133	95.68	95.00	95.34
1993	1	25.00	98.49	1.51	1	14	14	14	100.00	100.00	100.00
Total	88,541	46.88	97.48	2.52	231	3,962	3,832	3,645	95.12	92.00	93.53

Table 9. Validation results of the “Items Algorithm”

Notes: The table presents the validation results of the “Items Algorithm”. The second and third columns show the number of filings of which items could be extracted from by applying the “Items Algorithm”

(filings were not machine-parsable due to lacks of content, inconsistent filing structure, table tags and HTML formatting inconsistencies). Only filings with extracted items length exceeding 90 percent of 10-K section are presented. The next two columns show the average amount of extracted information from each filing in a particular year since 1993. The next columns show the performance evaluation of the “Items Algorithm” using precision (=number of correct answers/number of total answers), recall (=number of correct answers/total possible correct answers), and F-measure (=2*precision*recall/precision+recall).

7. DESCRIPTIVE STATISTICS ON FORM 10-K CONTENTS

In total, I examine the textual composition of 188,875 annual reports filed with the SEC between 1993 and 2016. On average, an annual report on Form 10-K submitted to the EDGAR system during the sample period is composed of 38,240 words. The average word count of an annual submission increased from 39,730 in 1994 to 46,111 in 2016. The medians of the word counts increased accordingly. The majority of textual information embedded in an annual report on Form 10-K are contained in the core document (64.95 percent) whereas the disclosed exhibits represent only a minority of the overall textual elements stated in annual submissions (35.04 percent). By examining the EDGAR database and its Form 10-K filings in more detail, investors and researchers can see that the average file size (Megabyte) of an annual report made with the electronic disclosure system increased in recent years due to HTML formatting, ASCIIencodings and XBRL documents. Table 10 presents descriptive statistics of the text length and the file size of 188,875 annual reports on Form 10-K (Form 10-K405) filed with the SEC between 1993 and 2016.

Table 10. Descriptive statistics of SEC EDGAR Form 10-K reports

Year	Filings	Word Count					File Size		
		Full Report (Number)			10-K Sections (%)	Exhibits (%)	Mean (MB)	Med. (MB)	Max. (MB)
		Mean	Med.	Max.					
2016	6,467	46,111	39,997	1,112,167	79.54	20.46	12.50	9.11	261.90
2015	7,985	43,909	37,262	1,657,009	79.51	20.49	15.12	10.18	414.52
2014	8,084	43,501	35,840	2,884,474	78.38	21.62	14.08	9.72	402.86
2013	8,105	43,884	35,181	6,257,121	77.32	22.68	13.22	9.38	254.18
2012	8,393	41,354	34,135	1,441,676	78.62	21.37	8.68	4.90	139.48
2011	8,840	41,087	33,008	1,031,964	77.33	22.67	4.48	1.71	212.57
2010	9,165	40,584	32,448	957,870	77.65	22.35	2.50	1.49	95.27
2009	9,839	40,406	32,074	3,997,528	74.97	25.03	1.90	1.33	86.21
2008	8,746	39,183	32,501	779,558	72.72	27.28	1.72	1.27	61.97
2007	8,574	39,761	32,206	2,617,579	73.67	26.33	1.81	1.28	91.99
2006	8,852	36,910	30,247	908,916	70.76	29.24	1.42	1.01	61.16
2005	9,017	36,166	28,854	1,442,810	66.13	33.86	1.19	0.82	80.62
2004	8,567	38,633	28,655	1,008,146	60.55	39.45	0.98	0.67	27.82
2003	8,468	39,193	28,738	911,982	58.15	41.83	0.90	0.55	24.01
2002	8,927	37,255	26,201	1,545,636	52.82	47.15	0.59	0.34	26.59
2001	9,248	35,153	24,531	1,308,749	52.03	47.97	0.40	0.28	23.34
2000	9,869	33,969	23,619	1,258,064	51.01	48.97	0.35	0.26	19.91
1999	10,122	33,634	23,290	496,458	49.40	50.56	0.33	0.25	8.29
1998	10,287	35,334	22,206	667,721	44.27	55.71	0.33	0.24	4.82
1997	9,899	32,269	20,496	650,347	44.84	55.14	0.30	0.22	4.82
1996	6,258	29,069	19,082	447,469	45.68	54.31	0.28	0.21	4.25
1995	3,236	34,803	22,570	361,832	38.51	61.48	0.34	0.24	4.03
1994	1,923	39,730	25,510	553,782	37.55	62.45	0.39	0.28	4.27
1993	4	20,571	18,247	31,993	83.01	16.99	0.23	0.26	0.27
Total	188,875	38,240	28,772	6,257,121	64.95	35.04	3.56	0.71	414.52

Notes: The table presents descriptive statistics of the text lengths, document compositions and file sizes for all annual reports filed with the SEC since 1993. Columns 3-5 show the means, medians and maxima of

word counts of Form 10-K filings made on EDGAR. The average distribution of textual information between the 10-K sections and exhibits contained in the “Complete Submission Text Files” is presented in column 6 and 7.

The distribution of textual elements among the various 10-K items is unequal. On average 22.65 percent of all textual information are contained in Item 1 (“Business”). Describing a company’s business as well as its main products and services, the item may also include information about the competition, regulations and other issues a particular company is faced with [46] [47]. Item 7 (“Management’s Discussion and Analysis of Financial Condition and Results of Operations – MD&A”) represents 18.58 percent of the given information within Form 10-K filings made with the SEC. The item states information about a company’s operations and financial results in addition to its liquidity and capital resources. The section may include off-balance sheet arrangements and contractual obligations alongside key business risks [46] [47]. Item 8 (“Financial Statements and Supplementary Data”) requires a company to disclose audited financial statements [46] [48] [47]. Additional information explaining the financial statements in more detail (“Notes to Consolidated Financial Statements”, “Report of Management”, “Report of Independent Registered Accounting Firm” etc.) represent 15.96 percent of all given information in the 10-K section of an annual report. Item 1A (“Risk Factors”) describes significant factors that may adversely affect a filer’s business, financial condition or future financial performance [46] [47]. Since electronic filings became available on average 8.42 percent of all textual information disclosed in annual submissions are contained in this section. Each of the remaining items only represent a fraction of the overall textual information embedded in Form 10-K filings. While the length for most sections in annual reports remained constant over time the amount of textual information contained in Item 1A (“Risk Factors”) increased from 12.56 percent in 2006 to 20.10 percent in 2016 indicating that SEC EDGAR filers disclose more information about risks in recent years.

8. SUMMARY

This paper displays the huge amount and variety of publicly available corporate information filed with the SEC and distributed by its EDGAR database. It shows how massive data can be retrieved from the SEC server in a fast and efficient way using simple and easy accessible software. The second main purpose of this paper is to create standardized procedures (“Annual Report Algorithm” and “Items Algorithm”) investors and researchers can use to extract any kind of textual information from financial statements filed with the SEC. This is achieved by providing regular expressions for multiple steps of data cleaning and filtering. Using these dynamic and platform-independent extraction algorithms the paper analyses the textual composition of more than 180,000 annual reports filed with the SEC via the EDGAR system between 1993 and 2016. The algorithms are tested for validity in several ways. The tools and algorithms intend to reduce costs and lower technical boundaries for researchers in the field of finance and accounting to engage in textual analysis.

ACKNOWLEDGEMENTS

I appreciate the valuable comments by Klaus Henselmann and Daniel Büchs

REFERENCES

- [1] Grant G H, Conlon S J (2006) EDGAR Extraction System: An Automated Approach to Analyze Employee Stock Option Disclosures, in: *Journal of Information Systems (JIS)*, 20(2)/2006, 119-142

- [2] Wilks Y (1997) Information Extraction as a Core Language Technology, in: M T Pazienza, Ed. Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer-Verlag, Berlin Heidelberg, Germany
- [3] Mooney R J, Bunescu R (2005) Mining Knowledge from Text Using Information Extraction, in: SIGKDD Explorations (SIGKDD), 7(1)/2005, 3-1
- [4] Gaizauskas R, Humphreys K, Azzam S, Wilks Y (1997) Concepticons vs. Lexicons: an Architecture for Multilingual Information Extraction, in: M T Pazienza, Ed. Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer-Verlag, Berlin Heidelberg, Germany
- [5] Garcia D, Norli O (2012) Crawling EDGAR, in: The Spanish Review of Financial Economics (SRFE), 10/2012, 1-10
- [6] Stümpert T (2008) Extracting Financial Data from SEC Filings for US GAAP Accountants, in: D Seese, C Weinhardt, F Schlottmann, Eds. Handbook on Information Technology in Finance. Springer-Verlag, Berlin Heidelberg, Germany
- [7] Bovee M, Kogan A, Nelson K, Srivastava R P, Vasarhelyi M A, (2005) Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL), in: Journal of Information Systems (JIS), 19(1)/2005, 19-41
- [8] O'Riain S (2012) Semantic Paths in Business Filings Analysis. Ph.D. thesis, National University of Ireland, Galway, Ireland
- [9] Loughran T, McDonald B (2014) Measuring Readability in Financial Disclosures, in: The Journal of Finance (JoF), 69(4)/2014, 1643-1671
- [10] Gerdes J Jr (2003) EDGAR-Analyzer: automating the analysis of corporate data contained in the SECs EDGAR database, in: Decision Support Systems 35/2003, 7-29
- [11] Kambil A, Ginsburg M (1998) Public Access Web Information Systems: Lessons from the Internet EDGAR Project, in: Communications of the ACM (CACM), 41(7)/1998, 91-97
- [12] Davis A K, Tama-Sweet I (2012) Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A, in: Contemporary Accounting Research (CAR), 29(3)/2012, 804-837
- [13] Loughran T, McDonald B (2016) Textual Analysis in Accounting and Finance: A Survey, in: Journal of Accounting Research (JAR), 54(4)/2016, 1187-1230
- [14] Tetlock P C (2007) Giving Content to Investor Sentiment: The Role of the Media in the Stock Market, in: The Journal of Finance (Jof), 62(3)/2007, 1139-1168
- [15] Loughran T, McDonald B (2011a) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, in: The Journal of Finance (JoF), 66(1)/2011, 35-65
- [16] Jegadeesh N, Wu D (2013) Word power: A new approach for content analysis, in: Journal of Financial Economics (JFE), 110(3)/2013, 712-729
- [17] Stümpert T, Seese D, Centinkaya Ö, Spöth R (2004) EASE – a software agent that extracts financial data from the SEC's EDGAR database, in: Proceedings of the 4th International ICSC Symposium on Engineering of Intelligent Systems (EIS 2004). Funchal, Portugal
- [18] Engelberg J, Sankaraguruswamy S (2007) How to Gather Data Using a Web Crawler: An Application Using SAS to Search Edgar. Working Paper, SSRN
- [19] Cong Y, Kogan A, Vasarhelyi M A (2007) Extraction of Structure and Content from the Edgar Database: A Template-Based Approach, in: Journal of Emerging Technologies in Accounting (JETA) 4(1)/2007, 69-86
- [20] Thai V, Davis B, O'Riain S, O'Sullivan D, Handschuh S (2008) Semantically Enhanced Passage Retrieval for Business Analysis Activity, in: Proceedings of the 16th European Conference on Information Systems (ECIS 2008). Galway, Ireland
- [21] Chakraborty V, Vasarhelyi M A (2010) Automating the process of taxonomy creation and comparison of taxonomy structures, in: 19th Annual Research Workshop on Strategic and Emerging Technologies, American Accounting Association. San Francisco, California, USA
- [22] Hernandez M A, Ho H, Koutrika G, Krishnamurthy R, Popa L, Stanoi I R, Vaithyanathan S, Das S (2010) Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance. IBM Technical Report #RJ10475
- [23] Srivastava R P (2016) Textual Analysis and Business Intelligence in Big Data Environment: Search Engine versus XBRL, in: Indian Accounting Review (IAR), 20(1)/2016, 1-20
- [24] SEC (2013) What We Do, Available online on URL: <https://www.sec.gov/about/whatwedo.shtml>
- [25] SEC (2010) Important Information about EDGAR, Available online on URL:

- <https://www.sec.gov/edgar/aboutedgar.htm>
- [26] SEC (2006) Electronic Filing and the EDGAR System: A Regulatory Overview, Available online on URL <https://www.sec.gov/info/edgar/regoverview.htm>
- [27] Pagell R A (1995) EDGAR: Electronic Data Gathering and Receiving, in: Business Information Review (BIR), 11(3)/1995, 56-68
- [28] SEC Release 34-36997 (1996) EDGAR Phase-in Complete on May 6, 1996, Available online on URL: <https://www.sec.gov/info/edgar/ednews/34-36997.htm>
- [29] SEC Regulation S-T (2016) General Rules and Regulations for electronic Filings, Available online on URL: http://www.ecfr.gov/cgi-bin/textidx?node=17:3.0.1.1.14&rgn=div5#se17.3.232_1100
- [30] SEC Release 33-8099 (2002) Mandated EDGAR Filing for Foreign Issuers, Available online on URL: <https://www.sec.gov/rules/final/33-8099.htm>
- [31] SEC (2015) Information for FTP Users, Available online on URL: <https://www.sec.gov/edgar/searchedgar/ftpusers.htm>
- [32] SEC Index Files (2016) Full Index Files, Available online on URL: <ftp://ftp.sec.gov/edgar/fullindex/>
- [33] W3C Recommendation (1999) HTML 4.01 Specification, Available online on URL: <https://www.w3.org/TR/html401/cover.html>
- [34] Filer Manual (2016) Filer Manual – Volume II EDGAR Filing, Available online on URL: <https://www.sec.gov/info/edgar/edgarfm-vol2-v37.pdf>
- [35] Ditter D, Henselmann K, Scherr E (2011) Using XBRL Technology to Extract Competitive Information from Financial Statements, in: Journal of Intelligence Studies in Business (JISIB), 1/2011, 19-28
- [36] W3C Strict DTD (1999) HTML 4.01 Strict DTD, Available online on URL: <https://www.w3.org/TR/html4/strict.dtd>
- [37] SEC (2000) HTML Specifications for EDGAR Rel. 7.0, Available online on URL: <https://www.sec.gov/info/edgar/ednews/edhtml.htm> [38] Bodnaruk A, Loughran T, McDonald B (2015) Using 10-K Text to Gauge Financial Constraints, in: Journal of Financial and Quantitative Analysis (JFQA), 50(4)/2015, 623-646
- [39] Loughran T, McDonald B (2011b) Internet Appendix for “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks”, Available online on URL: <http://www.afajof.org/SpringboardWebApp/userfiles/afa/file/Supplements%20and%20Data%20Sets/Internet%20Appendix%20for%20When%20Is%20a%20Liability%20Not%20a%20Liability%20Textual%20Analysis,%20Dictionaries,%20and%2010-Ks%206989-IA-Feb-2011.pdf>
- [40] SEC EDGAR Archives (2016) Coca Cola Company’s Financial Statement Submissions on 2016-02-25, Available online on URL: <https://www.sec.gov/Archives/edgar/data/21344/000002134416000050/0000021344-16-000050-index.htm>
- [41] Friedl J E F (2006) Mastering Regular Expressions, Third Edition, O’Reilly Media, Inc., Sebastopol, California, USA
- [42] Thompson K (1968) Regular Expression Search Algorithm, in: Communications of the ACM (CACM), 11(6)/1968, 419-422
- [43] Goyvaerts J, Levithan S (2012) Regular Expressions Cookbook, Second Edition, O’Reilly Media, Inc., Sebastopol, California, USA
- [45] SEC EDGAR (2015) Public Dissemination Service (PDS) Technical Specification, Available online on URL: https://www.sec.gov/info/edgar/specifications/pds_dissemination_spec.pdf
- [46] SEC (2011) Fast Answers – How to Read a 10-K, Available online on URL: <https://www.sec.gov/answers/reada10k.htm>
- [47] SEC Regulation S-K (2016) Standard Instructions for filing Forms under Securities Act of 1933, Securities Exchange Act of 1934 and Energy Policy and Conservation Act of 1975-Regulation S-K, Available online on URL: http://www.ecfr.gov/cgi-bin/textidx?SID=8e0ed509ccc65e983f9eca72ceb26753&node=17:3.0.1.1.11&rgn=div5#se17.3.229_1101
- [48] SEC Regulation S-X (2016) Form and Content of and Requirements for Financial Statements; Securities Act of 1933, Securities Exchange Act of 1934, Investment Company Act of 1940, Investments Advisers Act of 1940, and Energy Policy and Conservation Act of 1975- Regulation S-X, Available online on URL: http://www.ecfr.gov/cgi-bin/textidx?SID=8e0ed509ccc65e983f9eca72ceb26753&node=17:3.0.1.1.8&rgn=div5#se17.3.210_11_601

A STUDY ON THE MOTION CHANGE UNDER LOADED CONDITION INDUCED BY VIBRATION STIMULATION ON BICEPS BRACHII

Koki Honda¹ and Kazuo Kiguchi²

^{1,2} Department of Mechanical Engineering, Kyushu University, Fukuoka, Japan

ABSTRACT

To assist not only motor function but also perception ability of elderly and/or handicapped persons, the power-assist robots which have perception-assist function have been developed. These robots can automatically modify the user's motion when the robot detects inappropriate user's motion or a possibility of accident such as collision between the user and obstacles. For this motion modification in perception-assist, some actuators of power-assist robot are used. On the other hand, since some elderly persons, handicapped persons or some workers need not use power-assist function but perception-assist function only, another new concept perception-assist method was investigated in our previous study. In this perception-assist method, only vibrators are used for generating motion change with kinesthetic illusion to assist perception-ability only. In this study, since the perception-assist is often used during tasks under a loaded condition, the features of motion change under the loaded condition are investigated.

KEYWORDS

Vibration Stimulation, Perception-Assist, Motion Change, Elbow Joint Motion

1. INTRODUCTION

Since not only motor function but also perception ability to surrounding environment are deteriorated in some elderly and/or handicapped persons, some power-assist exoskeleton robots which have perception-assist function have been proposed [1], [2]. The power-assist robot with perception-assist function keeps monitoring the interaction between the user and the environment by using its sensors and if the user is making inappropriate motion or dangerous motion, the robot modifies his/her motion automatically by generating additional external force with some actuators.

On the other hand, there are also some elderly persons, handicapped persons and workers whose motor function is not so deteriorated but only perception ability has problem. To assist such persons, another new concept perception-assist method has been proposed in our previous study [3], [4]. In this perception-assist method, a vibrator is used to generate motion change with kinesthetic illusion and it has a possibility to realize an effective perception-assist device.

When vibration stimulation is added to the tendon of antagonist muscle around a human joint, a person feel as if their antagonist is elongated and feel as if their joint is rotating. This

phenomenon is called as “kinaesthetic illusion” and it is discovered by Goodwin in 1972 [5]. This kinaesthetic illusion has been studied in the field of neurophysiology. The cause of this illusion is that the receptors in a muscle, called as muscle spindle, is stimulated by vibration and it generate electric signals, called as Ia afferents, to brain. By receiving this Ia afferents, the brain makes misinterpretation that the vibrated muscle is elongated and the body parts which is connected to vibrated muscle are moving despite the body parts are not moving actually [6], [7].

We found that the motion change can be generated by using this misinterpretation generated by vibration stimulation in previous studies [3], [4]. In those studies, vibration stimulation is added to biceps brachii and triceps brachii during elbow joint flexion/extension motion and owing to the gap between actual elbow joint angular velocity and subject’s elbow joint feeling, motion change was generated.

In this study, vibration stimulation is added to biceps brachii during elbow extension motion under some loaded conditions. Since many daily tasks are conducted under a loaded condition, the features of motion change under the loaded condition induced by vibration stimulation must be investigated. The experimental results show that the motion change can be generated under the loaded condition by vibration stimulation and the changing rates of motion change are investigated.

2. METHODS

2.1. Experimental devices

In this study, vibration stimulation is added to biceps brachii of subject’s right elbow. Subjects are seated on a chair (see Figs. 1 and 2). Subject’s both arms are fixed to the frame of goniometer (rotary encoder: RE12D-300-201). The angle range of elbow extension motion is 0 (deg)-45 (deg). Elbow angle and angular velocity are recorded by workstation. According to previous study [3], 70-100 Hz vibration stimulation can generate kinaesthetic illusion and motion change. The vibrator used in this study (Fig.1) can also generate 70-100 (Hz) vibration stimulation and its amplitude is 1.0 (mm).

2.2. Procedure of the experiments

In this experiment, the effect of loaded condition to motion change caused by vibration stimulation is investigated. To make the amount of motion change with kinaesthetic illusion clear, subjects are assigned to make the elbow angle of the right “Vibrated arm” correspond to the elbow angle of left “Reference arm”. It is assumed that a gap between the elbow angle of Vibrated arm and Reference arm is generated when an illusion is generated with vibration stimulation. Procedure of the experiment is shown below. There are a practice part and an experiment part.

A) Practice part

1. Practices are conducted to position their elbow angle of both arm 20 (deg) without watching.
2. Other practices are conducted to generate extension motion with both elbow angles’ angular velocity 3 (deg/s).

B) Experiment part

1. Subject's elbow angle is adjusted to 0 (deg).
2. The encoders recording are started and subject's both elbow angles are adjusted to 45 (deg).

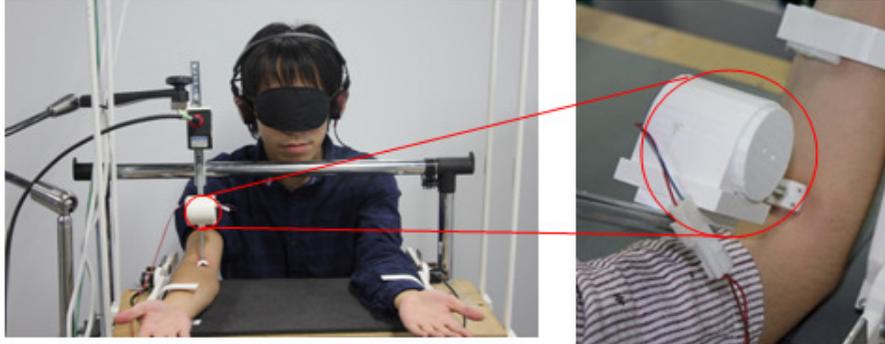


Figure 1. Experimental device and vibrator

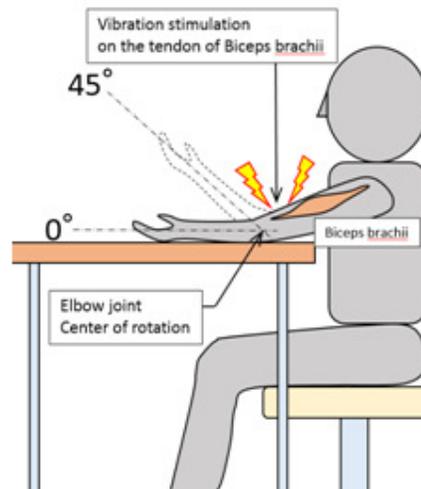


Figure 2. Experimental conditions

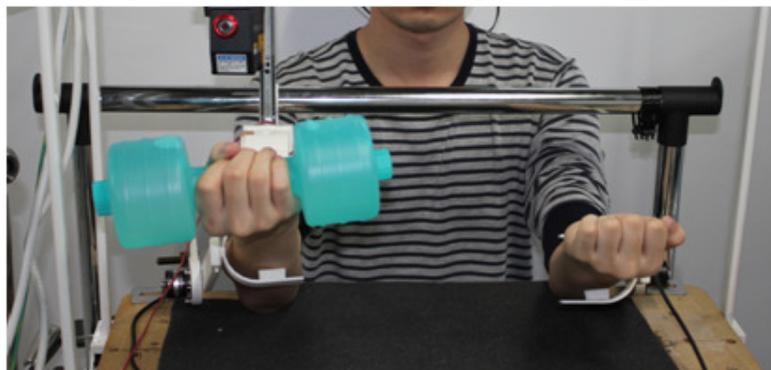


Figure 3. Loaded condition during motion change

3. Experimenter gives a weight to the subject and the subject hold the weight by his/her hand (see Fig. 3). After that, vibration stimulation is started and then extension motion with 3 (deg/s) of both arms are started 10 seconds later.
4. When the reference arm comes up to 45(deg), encoders are turned off and vibration stimulation is ended.
5. Once trial is finished, a weight is changed by the experimenter and the next procedure is restarted.

Subjects are shown as bellow (Table 1). We have 3 patterns of loaded conditions; 1000 (g), 500 (g), and 0 (g). Subject's eyes are closed and subjects hear white noise during the experiment.

Table 1. Subjects

	Sex	Age	Weight (kg)	Height (cm)
Subject1	Man	23	57	166
Subject2	Man	25	57	156
Subject3	Man	23	62	178

3. RESULTS

Trials are conducted 3 times on each loaded conditions with each subject. From the records of subject's elbow joint angle during the experiments, the changing rate of the gap between the Reference arm and the Vibrated arm is calculated (see Fig. 4). In almost all subjects, the data of the elbow angles of both arms during 13s to 18s, after vibration stimulation start, are used in this calculation. The average of the changing rates in 3 trials and elbow angle of the Vibrated arm are shown in Table 2.

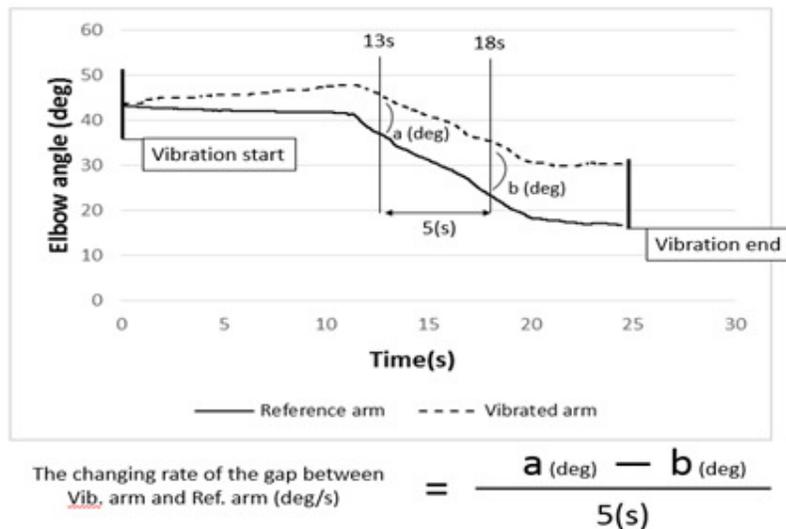


Figure 4. Calculation Method

Table 2 Results of experiments

		0 (g)	500 (g)	1000 (g)
Subject1	Changing rate(deg/s)	0.8	1.2	1.5
	Angle (deg)	45	40	38
Subject2	Changing rate(deg/s)	1.6	1.2	1.4
	Angle (deg)	42	36	44
Subject3	Changing rate(deg/s)	2.4	3.1	1.6
	Angle (deg)	44	46	48

4. DISCUSSION

In all subjects, the gap between the Vibrated arm's elbow angle and the Reference arm's elbow angle was generated. These experimental results suggest that motion change can be generated under loaded condition by vibration stimulation.

The correlation between the increasing of load and the increasing or decreasing of changing rate was not seen clearly in these experiments. There is a possibility that the kinaesthetic illusion is enhanced as the load increases since the muscle spindles are more strongly elongated by load. On the other hand, as the angle of the elbow joint decreases, however, the moment due to the load applied to the elbow joint increases because of the gravity effect. Consequently, there is a possibility that the gap between the Vibrated arm's elbow angle and the Reference arm's elbow angle was decreased by the effect of the moment.

The results suggests that there is a possibility that the correlation between the load increase and the increase/decrease of the changing rate can be obtained more clearly if the constant torque is added to the Vibrated arm's elbow joint during its extension motion.

ACKNOWLEDGEMENTS

This work was partially supported by Japan Society of Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) 16H04305.

REFERENCES

- [1] K. Kiguchi, M. Liyanage, and Y. Kose (2009) "Perception Assist with an Active Stereo Camera for an Upper-Limb Power-Assist Exoskeleton", *Int. Journal of Robotics and Mechatronics*, vol. 21, no. 5, pp.614-620.
- [2] K. Kiguchi and Y. Hayashi (2011) "A Lower-Limb Power-Assist Robot with Perception-Assist", *Proc. of IEEE Int. Conf. on Rehabilitation Robotics*, pp.731-736.

- [3] Koki Honda, Kazuo Kiguchi (2016) "A Fundamental Study on the Effect of Vibration Stimulation for Motion Modification in Perception-Assist", Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics.
- [4] Koki Honda, Kazuo Kiguchi (2016) "A Fundamental Study on the Effect of Vibration Stimulation on Triceps Brachii during Elbow Flexion motion for Perception-Assist", Proc. of Int. Symp. on Micro-Nanomechatronics and Human Science.
- [5] G. M. Goodwin, D. I. McCloskey, P. B. C. Matthews (1972) "The contribution of muscle afferents to kinaesthesia shown by vibration induced illusions of movement and by the effects of paralysing joint afferents." Brain, Vol.95, No.4, pp.705-748.
- [6] E. Naito, H.H. Ehrsson, S. Geyer, K. Zilles, and P.E. Roland (1999) "Illusory arm movements activate cortical motor areas: a positron emission tomography study", The Journal of neuroscience, vol.19, no.14, pp.6134-6144.
- [7] J.P. Roll and J.P. Vedel (1982) "Kinaesthetic role of muscle afferents in man, studied by tendon vibration and microneurography", Experimental Brain Research, vol.47, no.2, pp.177-190.

AUTHORS

Koki Honda received B. Eng. Degree in mechanical engineering from Kyushu University, Fukuoka, Japan, in 2015, He is currently master course student of Kyushu University, Fukuoka, Japan.



Kazuo Kiguchi received the B. Eng. degree in mechanical engineering from Niigata University, Niigata, Japan, in 1986, the M. A Sc. degree in mechanical engineering from the University of Ottawa, Ottawa, Canada, in 1993, and the Dr. Eng. degree from Nagoya University, Nagoya, Japan, in 1997. He is currently a professor in the Dept. of Mechanical Engineering, Faculty of Engineering, Kyushu University, Japan.



ADAPTIVE AUTOMATA FOR GRAMMAR BASED TEXT COMPRESSION

Newton Kiyotaka Miura¹ and João José Neto¹

¹Escola Politécnica da Universidade de São Paulo, São Paulo, Brazil

ABSTRACT

The Internet and the ubiquitous presence of computing devices anywhere is generating a continuously growing amount of information. However, the information entropy is not uniform. It allows the use of data compression algorithms to reduce the demand for more powerful processors and larger data storage equipment. This paper presents an adaptive rule-driven device - the adaptive automata - as the device to identify repetitive patterns to be compressed in a grammar based lossless data compression scheme.

KEYWORDS

Adaptive Automata, Grammar Based Data Compression

1. INTRODUCTION

The amount of data processed by the computers has grown by the increase in hardware computing power and the data storage capacity. New applications are built to take advantage of this, reinforcing the need for more capacity in data processing. Currently, analysis of data generated by social media Internet applications and genome database processing are examples of applications that require handling of huge amount of data. In this scenario, the necessity for optimizing the use of the finite amount of data storage available in computers is economically justifiable.

Grammar-based text compression is a method for representing a sequence of symbols using a context-free grammar (CFG) as defined in Chomsky's hierarchy [1], which generates a single sequence identical to the original one. It is based on the idea that a CFG can compactly represent the repetitive structures within a text. Intuitively, greater compression would be obtained for input strings that contain a greater number of repeated substrings that will consequently be represented by the same grammatical production rule. Examples of data with these characteristics are the sequences of genes of the same species, texts with version control systems.

The naturally hierarchical definition of a CFG allows string-manipulation algorithms to perform operations directly on their compressed representations, without the need for a prior decompression [2] [3] [4] [5]. It potentially brings the advantage of decreasing the temporary storage space required for data manipulation, in addition to opening the possibility of the algorithms present shorter execution times by decreasing the data to be processed [2]. These features are attractive for improving the efficiency in processing large amounts of data whose demand has grown mainly in Big Data applications and manipulation of genomes.

Operations in grammatically compressed texts [6] include string search, edit distance calculation, string or character repeating frequency calculation, access to a specific position of the original string, obtaining the first occurrence of a character and indexing for random access. Examples of

grammar-based compression applications such as repetitive structure mining, pattern recognition, data mining using are cited in [7], [3] and [2].

The process of obtaining this type of grammar with specific characteristics is a grammatical inference process, which is studied within the computer sciences since the sixties [8].

This work focuses on adopting an adaptive device guided by rules [9] for the identification of repetitive data for grammar-based text compression.

An adaptive device [9] has the ability of self-modification, that is, it changes the rules of its operation at execution time without the need for external intervention. An example is the adaptive automaton, which is a state machine with computational power equivalent to the Turing machine [10]. It might change its configuration based on the input string.

In the Section 2 the basic concept of this type of compression and some algorithms found in the literature is presented. In the Section 3 an adaptive automaton-based implementation of the repetitive pattern searching is presented.

2. BACKGROUND

2.1. Grammar based compression

Text compression is performed by a CFG $G = (\Sigma, V, D, X_s)$, where Σ is the finite set of terminal symbols. V is the set of nonterminal (or variable) symbols with $\Sigma \cap V = \emptyset$. $D \subset V \times (V \cup \Sigma)^*$ is the finite set of rules of production with size $n = |V|$. $X_s \in V$ is the non-terminal that represents the initial nonterminal symbol of the grammar.

The syntax tree of G is an ordered binary tree in which the inner nodes are labelled with the non-terminal symbols of V and the leaves with the terminals of Σ , that is, the sequence of labels in the sheets Corresponds to the input string S .

Each internal node Z corresponds to a production rule $Z \rightarrow XY$ with the child nodes X on the right and Y on the left. As G can be expressed in the normal form of Chomsky [1] any compression grammar is a Straight Line Program (SLP) [11] [12] [2] which is defined as a grammatical compression on $\Sigma \cup V$ and production rules in the form $X_k \rightarrow X_i X_j$ where $X_k, X_i, X_j \in \Sigma \cup V$ and $1 \leq i, j < k \leq n + \sigma$.

The compression of a string S of the size $\sigma = |S|$ is achieved when the σ is greater than the sum of the size of the grammar G that generates it and the size of the compressed sequence.

Figure 1 presents an example of the compression of the string $s = (a, b, a, a, b, c, a, b, a, a)$, resulting in the compressed sequence $S = \{X_4, c, X_3\}$ and the derivation dictionary $D = \{X_1 \rightarrow ab, X_2 \rightarrow aa, X_3 \rightarrow X_1 X_2, X_4 \rightarrow X_3 X_1\}$, corresponding to the forest of syntactic trees. The dashed lines of the same colour identify portions of the tree that have parent nodes with the same label.

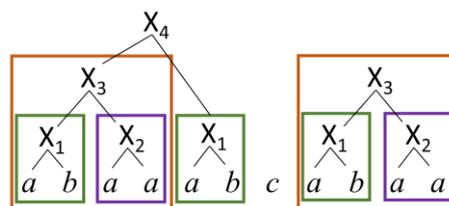


Figure 1 Example of grammar based compression.

The main challenge of grammar-based compression is to find the smallest CFG that generates the original string to maximize the compression rate. This problem has been shown to be intractable. Storer and Szymanski demonstrated [7] that given a string S and a constant k , obtaining a CFG of size k that generates S is an NP-complete problem. Furthermore, Charikar et al. [13] demonstrated that the minimum CFG can be approximated by a logarithmic rate calculating that $8569/8568$ is the limit of this approximation rate of the algorithms to obtain the lowest value of k , assuming that $P \neq NP$.

Much research effort has focused in finding approximate minimal grammar inference algorithms, considering not only the compression but also searching grammars and data structures to represent them with characteristics suitable to support operations in the compressed text [2] [3] [4] [5]. Regarding the dictionary representation, as discussed by [3] several initial algorithms have adopted Huffman coding to compact it although it does not allow random access of the strings, and more recently the succinct data structure method has been used. A brief description of some researches are presented below.

2.2. Compression algorithms

In the descriptions below N is the size of the input string, g is the minimum CFG size, and \log refers to \log_2 .

The Sequitur algorithm proposed by Nevill-Manning and Witten [14] operates incrementally in relation to the input string with the restriction that each bigram is present only once in a derivation rule of the inferred grammar and that each rule is used more than once. Sequitur operates in a linear space and execution time relative to the size of the input string.

The RePair algorithm developed by Larsson and Moffat [15] constructs a grammar by iteratively replacing pairs of symbols, either terminal or non-terminal, with a non-terminal symbol by doing an off-line processing of the complete text, or long phrases, and adopting a compact representation of the dictionary. It has the following simple heuristic to process a sequence S :

1. Identify the most frequent pair ab in S .
2. Add a rule $X \rightarrow ab$ to the dictionary of productions, where X is a new symbol that is not present in S .
3. Replace every occurrence of the pair ab in S by the new symbol X .
4. Repeat this procedure until any pair in S occurs just once.

Although it requires a memory space above 10 times the size of the input string, the algorithm presents a linear execution time, being efficient mainly in the decompression, facilitating search operations in the compressed data.

Rytter [16] and Charikar et al. [13] have developed algorithms that approximate the size of the CFG obtained in $O(\log N/g)$ by transforming the representation of the data with LZ77 method [17] to the CFG. Rytter [16] proposed the use of a grammar whose derivation tree is a balanced AVL binary tree in which the height of two daughter sub-trees differ only by one unit, which favours pattern matching operations. Charikar et al. [13] imposed the condition that the binary derivation tree be balanced in width, favouring operations such as union and division.

Sakamoto [18] adopted a strategy similar to RePair [15] and obtained an algorithm requiring a smaller memory space, performing iterative substitution of pairs of distinct symbols and repetitions of the same symbol, using double-linked lists for storing the input string and a priority queue for the frequency of the pairs.

Jez [19] modified the algorithm of [18] obtaining a CFG of size $O(g \log(N/g))$ for input strings with alphabet Σ that can be identified by numbers of $\{1, \dots, N^c\}$ for a constant c . Unlike the previous research, it was not based on Lempel-Ziv representation.

Maruyama et al. [5] developed an algorithm based on a context-dependent grammar subclass Σ -*sensitive* proposed to optimize the pattern matching operation on the compacted data.

Tabei et al. [4] has devised a scalable algorithm for least-squares partial regression based on a grammar-packed representation of high-dimensional matrices that allows quick access to rows and columns without the need for decompression. Compared to probabilistic techniques, this approach showed superiority in terms of precision, computational efficiency and interpretability of the relationship between data and tag variables and responses.

The text compression is also a focus of research such as the fully-online compression algorithm (FOLCA) proposed by Maruyama et al. [20] which infers partial parsing trees [16] whose inner nodes are traversed post-order and stored in a concise representation. For class $C = \{x_1, x_2, \dots, x_n\}$ of n objects, \log_n is the minimum of bits to represent any $x_i \in C$. If the representation method requires $n + (n)$ bits for any $x_i \in C$, the representation is called succinct [3]. They present experimental results proving the scalability of the algorithm in memory space and execution time in processing human genomes with high number of repetitive texts with presence of noise.

Fukunaga et al. [11] proposed an online algorithm approach to approximate frequent patterns of grammatically compressed data with less memory consumption compared to offline methods. *Edit-sensitive parsing* [21], which measures the similarity of two symbol strings by the edit distance, is used for comparison of grammars subtree.

2.3. Grammar inference using adaptive technology

An adaptive rule driven device has the self-modifying capability [9], that is, it changes the rules of its operation according to the input data at run time without the need for external intervention. An example is the adaptive automaton [22] which consists of a traditional automaton with the addition of an adaptive mechanism. This mechanism allows modifications in the configuration of the underlying traditional automaton by invoking adaptive functions which can change its set of rules, enabling the adaptive automaton to have a computational power equivalent to the Turing machine [10].

Grammar-based compression is a specific case of grammatical inference whose purpose is to learn grammar from information available in a language [8]. In this case the available information corresponds to the text to be compressed which is the only sentence of a given language. José Neto and Iwai [23] proposed the use of adaptive automaton to build a recognizer with ability to learn a regular language from the processing of positive and negative samples of strings belonging to that language. The recognizer is obtained from the agglutination of two adaptive automata. One constructs the prefix tree from the input strings and the other produces a suffix tree from the inverse sequence of the same string. Matsuno [24] implemented this algorithm based on adaptive automata, and presented an application of the Charikar algorithm [13] to obtain a CFG from samples of the language defined by this grammar.

In this paper, we applied the adaptive automaton to infer a grammar to generate a compressed version of a string.

3. GRAMMAR BASED COMPRESSION USING ADAPTIVE AUTOMATA

Our approach using adaptive automaton is inspired in the algorithm of RePair [15]. The automaton is used in the process of finding pairs of symbols to be substituted by a grammar production rule.

The adaptive automaton modifies the configuration of the traditional finite automaton, which is represented by a tuple (Q, Σ, P, q_0, F) . The meaning of the elements of the tuple, along with other elements used in this work are described in the Table 1 for quick reference. We adopted a notation used by Cereda et al. [25].

Modification in the underlying automaton occurs through adaptive functions to be executed either before or after the transitions, according to the consumed input string symbols. The adaptive function executed before the transition is represented by the character ‘.’ written after the function name (e.g. $\mathcal{A} \cdot$) and the function executed after the transition is represented by the same character written before the name (e.g. $\cdot \mathcal{B}$). They can modify the automaton configuration by performing elementary adaptive actions for searching, exclusion or insertion of rules. The adaptive functions use variables and generators to perform editing actions in the automaton. Variables are filled only once in the execution of the adaptive function. Generators are special types of variables, used to associate unambiguous names for each new state created in the automaton and are identified by the ‘*’ symbol, for example, g_1^*, g_2^* .

Table 1. List of elements

Element	Meaning
Q	set of states
$F \subset Q$	subset of accepting states
$q_0 \in Q$	initial state
Σ	input alphabet
D	set of adaptive functions
P	$P: D \cup \{\varepsilon\} \times Q \times \Sigma \mapsto Q \times \Sigma \cup \{\varepsilon\} \times D \cup \{\varepsilon\}$, mapping relation
$\sigma \in \Sigma$	any symbol of the alphabet
$\mathcal{A}(q, x)$	adaptive function $\mathcal{A} \in D$ with arguments q, x triggered before a symbol consumption
$\mathcal{B}(y, z)$	adaptive function $\mathcal{B} \in D$ with arguments y, z triggered after the symbol consumption
g_i^*	generator used in adaptive functions that associates names with newly created states
$-(q_i, \sigma) \rightarrow (q_j)$	elementary adaptive action that removes the transition from q_i to q_j and consumes σ
$+(q_i, \sigma) \rightarrow (g_i^*, \varepsilon), \mathcal{A}$	rule-inserting elementary adaptive action that adds a transition from state q_i , consuming the symbol σ , and leading to a newly created state g_i^* with the adaptive function \mathcal{A} to be executed before the consumption of σ
Out_i	semantic action to be performed in state q_i , as in the Moore machine [1]

The adaptive automaton presented in this paper analyses trigrams contained in the original input string to search the most appropriate pair of symbols to be replaced by a grammar

production rule as in an iteration of the RePair [15] algorithm. From the initial state to any final state it will have a maximum of 3 transitions, as it analyses only 3 symbols. Thus, for each trigram the automaton restarts its operation from the initial state q_0 . This requires obtaining the set of all the trigrams present in the input string. For example, considering the sequence *abcab*, the input will be the set of trigrams $\{abc, bca, cab\}$.

The automaton counts the occurrence of every trigram and its containing prefix and suffix bigrams. This is accomplished by counters that are incremented by the semantic action functions Out_i executed in the states in which a bigram or trigram is recognized. Considering the trigram *abc*, it counts the occurrence of the trigram itself and the occurrence of *ab* and *bc* as a prefix and suffix.

Each bigram will have 2 counters, one for prefix and one for suffix, which can be incremented in any state of the automaton that has the associated semantic action function. Based on these counters, after processing all the trigrams it will be possible to choose the pair to be replaced. The pair that occurs most frequently inside the most frequent trigram, or the prefix or suffix bigram of this trigram are examples of criteria for the pair selection.

Starting from the terminal state of the automaton in which the most frequent trigram is recognized, it is possible to identify the constituent bigrams and get the values of their counters by traversing the automaton in the opposite direction of the transactions, that is, towards the starting state. In the bigram counter associated with a terminal state, it is obtained the total count of occurrences of the suffix bigram, and in the previous state the count of the prefix bigram.

The use of adaptive technique guides the design of this automaton by considering how it should be incrementally build as the input symbols are consumed, performing the actions just presented above.

In the following lines, we describe the processing of the trigram $\sigma_0\sigma_1\sigma_2$ of a hypothetical input string to better illustrate the configuration evolution of the automaton.

Figure 2 shows a simplified version of the starting configuration of the automaton composed by a single state q_0 and transitions for each symbol $\sigma \in \Sigma$ by executing the adaptive function $\mathcal{A}_0(\sigma, q_0)$ before its consumption. For ease of visualization, the representation of the set of these transitions has been replaced by a single arc in which the symbol to be consumed is indicated as $\forall\sigma$. This same simplification of representation was adopted in the description of the algorithm 1 (Figure 3) of the adaptive function \mathcal{A}_0 .

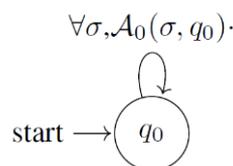


Figure 2. Initial topology of the adaptive automaton

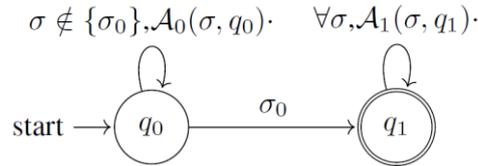
Function \mathcal{A}_0 modifies the automaton by removing the transition that consumes the first symbol σ_0 of the trigram, and creating three new elements: the state q_1 , the transition from q_0 to q_1 to allow the consumption of σ_0 and the loop transition from q_1 associating it with the consumption of $\forall\sigma \in \Sigma$ and another adaptive function \mathcal{A}_1 .

Algorithm 1: Adaptive function \mathcal{A}_0

adaptive function $\mathcal{A}_0(s, q_x)$
Generators: g_1^*
 $-(q_x, s) \rightarrow q_x$
 $+(q_x, s) \rightarrow (g_1^*, \epsilon)$
 $+(g_1^*, \forall \sigma) \rightarrow (g_1^*, \epsilon), \mathcal{A}_1$
end

Figure 3. Algorithm 1: Adaptive function \mathcal{A}_0

Figure 4 shows the new adaptive automaton topology after consumption of the first symbol σ_0 .

Figure 4. Topology after consuming the first symbol σ_0

The algorithm 2 (Figure 5) presents the adaptive function \mathcal{A}_1 . Its operation is similar to \mathcal{A}_0 creating a new state q_2 . It also prepares the consumption of a third symbol by inserting a transition with the adaptive function \mathcal{A}_2 , which is described in the algorithm 3 (Figure 6).

Algorithm 2: Adaptive function \mathcal{A}_1

adaptive function $\mathcal{A}_1(s, q_x)$
Generators: g_1^*
 $-(q_x, s) \rightarrow q_x$
 $+(q_x, s) \rightarrow (g_1^*, \epsilon)$,
 $+(g_1^*, \sigma) \rightarrow (g_1^*, \epsilon), \mathcal{A}_2$
end

Figure 5. Algorithm 2: Adaptive function \mathcal{A}_1

Function \mathcal{A}_2 (Figure 6) modifies the automaton configuration by removing the transition from q_2 to itself by consuming the symbol σ_3 (the third symbol of the trigram) creating a new state q_3 and the transition to it consuming σ_3 . The newly created state q_3 will be associated with the semantic function Out_1 .

Algorithm 3: Adaptive function \mathcal{A}_2

adaptive function $\mathcal{A}_2(s, q_x)$
Generators: g_1^*
 $-(q_x, s) \rightarrow q_x$
 $+(q_x, s) \rightarrow (g_1^*, \epsilon)$
end

Figure 6. Algorithm 3: Adaptive function \mathcal{A}_2

Figure 7 shows the automaton topology after consumption of the second and third symbol of the input string.

States q_2 and q_3 are associated with output functions Out_0 and Out_1 respectively. They are related to semantic actions in these states.

Out_0 is the function responsible for incrementing the occurrence counter of the prefix bigram $\sigma_0\sigma_1$. Out_1 is responsible for incrementing the occurrence counter of the suffix bigram $\sigma_1\sigma_2$, and the counter of the trigram $\sigma_0\sigma_1\sigma_2$.

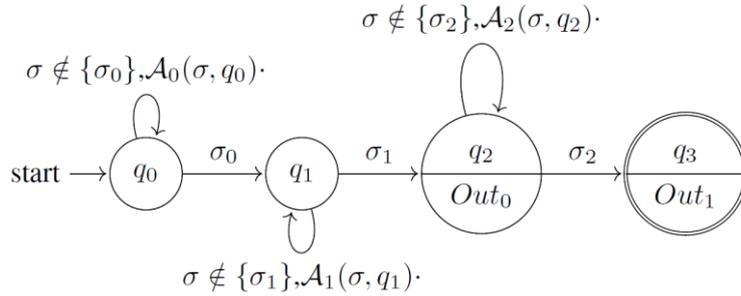


Figure 7. Topology after consuming the initial trigram $\sigma_0\sigma_1\sigma_2$

To better illustrate the operation of the adaptive automaton, Figure 8 shows its configuration after processing the sample string $abcaba$. The index i of the states q_i corresponds to the sequence in which they were entered by the algorithm.

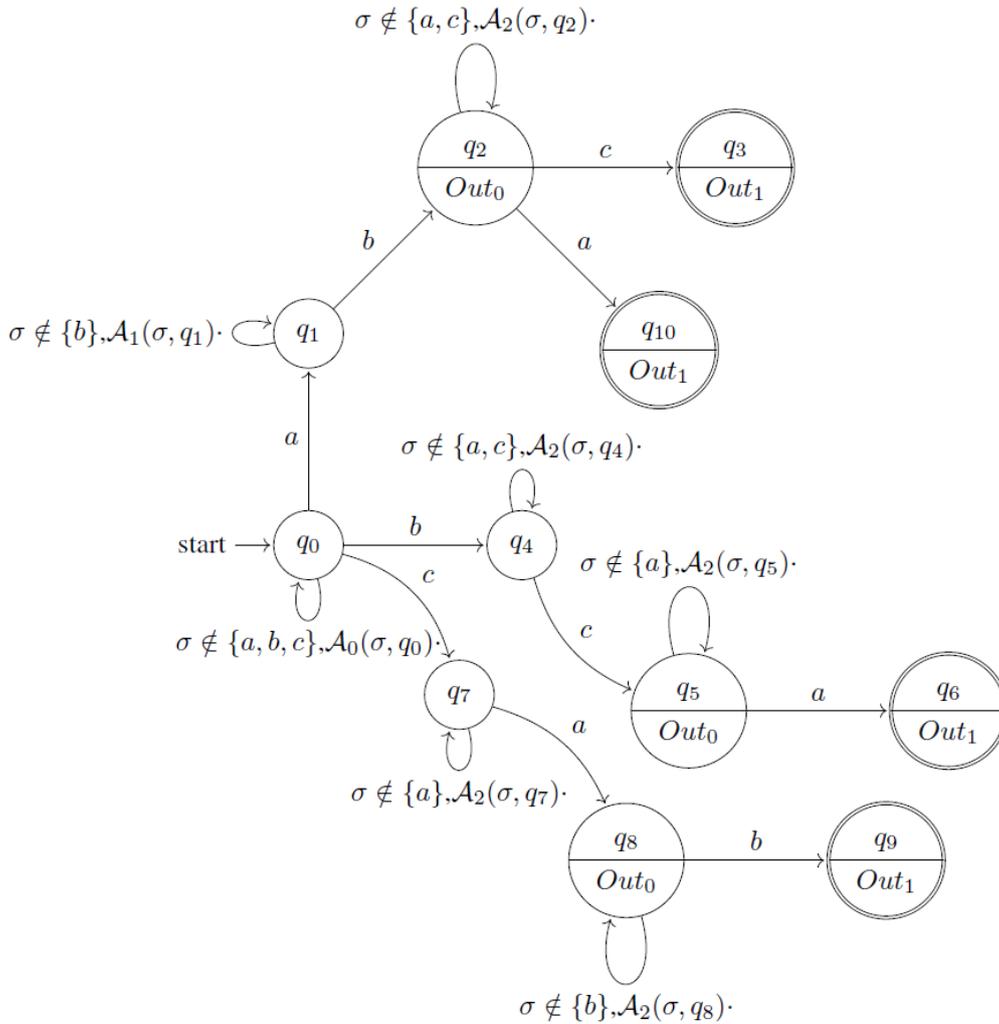


Figure 8. Automaton topology after processing $abcaba$

3. EXPERIMENTS

A test system is being developed using a Java language library for adaptive automaton¹ proposed by [26] as the core engine, with the surrounding subsystems such as the semantic functions and the iterations that substitutes the frequently occurring patterns.

Publicly available corpora² will be used as the input data.

Compression rates, execution time and temporary storage requirements for the different criteria of choice of repetitive patterns that the algorithm allows will be measured and compared to other techniques.

4. CONCLUSION

In this work, we presented the adaptive automaton as a device to identify a bigram, prefix or suffix, which are most repeated within the most repeated trigram in a sequence of symbols to obtain a substitution rule in a compression algorithm based on grammar.

In this paper, we presented the adaptive automaton as the device to identify repetitive bigrams in a sequence of symbols considering its presence inside frequently occurring trigrams. This bigram can be specified to be whether prefix or suffix of the trigram allowing the adoption of different configurations to be experimented.

As a future work, the adaptive automaton can be further expanded to analyse n-grams larger than trigrams. In addition, a comparative performance study could be done with other techniques. Another point to be analysed is the adoption of an adaptive grammatical formalism [27] in the description of the inferred grammar with the aim of making some operation in the compressed data.

Adaptive rule-driven device allows the construction of a large and complex system by simplifying the representation of the problem. This type of automaton is designed by specifying how the device must be incrementally modified in response to the input data, from a simple initial configuration, aiming at its desired intermediate configurations, and the output to be obtained by associated semantic actions.

ACKNOWLEDGEMENT

The first author is supported by Olos Tecnologia e Sistemas.

REFERENCES

- [1] Sipser, M. (2006) "Introduction to the theory of computation", Thomson Course Technology.
- [2] Lohrey, M. (2012) "Algorithmics on SLP-compressed strings: A survey." *Groups Complexity Cryptology*, vol. 4, no. 2, pp. 241–299.
- [3] Sakamoto, H. (2014) "Grammar compression: Grammatical inference by compression and its application to real data", in *Proceedings of the 12th International Conference on Grammatical Inference, ICGI 2014, Kyoto, Japan, September 17-19, 2014.*, pp. 3–20.
- [4] Tabei, Y., Saigo, H., Yamanishi, Y. & Puglisi, S. J. (2016) "Scalable partial least squares regression on grammar-compressed data matrices", in *22nd KDD*, pp. 1875—1884.

¹ <https://github.com/cereda/aa>

² <http://pizzachili.dcc.uchile.cl/repcorpus.html>

- [5] Maruyama, S., Tanaka, Y., Sakamoto, H., & Takeda, M. (2010) “Context-sensitive grammar transform: Compression and pattern matching”, *IEICE Transactions on Information and Systems*, vol. E93.D, no. 2, pp. 219–226.
- [6] Tabei, Y. (2016) “Recent development of grammar compression”, *Information Processing Society of Japan Magazine*, vol. 57, no. 2, pp. 172–178.
- [7] Jez, A. (2016) “A really simple approximation of smallest grammar”, *Theoretical Computer Science*, vol. 616, pp. 141 – 150.
- [8] De la Higuera, C. (2010) *Grammatical inference: learning automata and grammars*. Cambridge University Press.
- [9] José Neto, J. (2001) “Adaptive rule-driven devices - general formulation and case study”, in *CIAA 2001 6th International Conference on Implementation and Application of Automata*, ser. *Lecture Notes in Computer Science*, B. W. Watson and D. Wood, Eds., vol. 2494. Pretoria, South Africa: Springer-Verlag, pp. 234–250.
- [10] Rocha, R. L. A. & José Neto, J. (2000) “Adaptive automaton, limits and complexity compared to the Turing machine”, in *Proceedings of the I LAPTEC*, pp. 33–48.
- [11] Fukunaga, S., Takabatake, Y., I, T. & Sakamoto, H. (2016) “Online grammar compression for frequent pattern discovery”, *CoRR*, vol. abs/1607.04446.
- [12] Takabatake, Y., Tabei, Y., & Sakamoto, H. (2015) *Online Self-Indexed Grammar Compression*. Springer International Publishing, pp. 258–269.
- [13] Charikar, M., Lehman, E., Liu, D., Panigrahy, R., Prabhakaran, M., Sahai, A. & Shelat, A. (2005) “The smallest grammar problem.” *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2554–2576.
- [14] Nevill-Manning, C. G. & Witten, I. H. (1997) “Identifying hierarchical structure in sequences: A linear-time algorithm”, *J. Artif. Intell. Res.(JAIR)* vol. 7, pp. 67–82.
- [15] Larsson, N. J. & Moffat, A. (1999) “Offline dictionary-based compression”, in *Data Compression Conference, 1999. Proceedings. DCC '99*, pp. 296–305.
- [16] Rytter, W. (2002) *Application of Factorization to the Approximation of Grammar-Based Compression*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 20–31.
- [17] Ziv, J. & Lempel, A. (2006) “A universal algorithm for sequential data compression”, *IEEE Trans. Inf. Theor.*, vol. 23, no. 3, pp. 337–343.
- [18] Sakamoto, H., (2005) “A fully linear-time approximation algorithm for grammar-based compression”, *Journal of Discrete Algorithms*, vol. 3, no. 2–4, pp. 416–430.
- [19] Jez, A. (2015) “Approximation of grammar-based compression via recompression”, *Theoretical Computer Science*, vol. 592, pp. 115–134.
- [20] Maruyama, S. & Tabei, Y. (2014) “Fully online grammar compression in constant space.” in *DCC*, Bilgin, A., Marcellin, M. W., Serra-Sagristà, J. & Storer, J. A. Eds. IEEE, pp. 173–182.
- [21] Cormode, G. & Muthukrishnan, S. (2007) “The string edit distance matching problem with moves”, *ACM Transactions on Algorithms (TALG)* vol. 3, no. 1, pp. 2:1–2:19.
- [22] José Neto, J. (1994) “Adaptive automata for context-sensitive languages”, *SIGPLAN Notices*, vol. 29, no. 9, pp. 115–124.
- [23] José Neto, J. & Iwai, M. K. (1998) “Adaptive automata for syntax learning” in *Anais da XXIV Conferência Latinoamericana de Informática - CLEI 98*, pp. 135–149.
- [24] Matsuno, I. P. (2006) “Um estudo dos processos de inferência de gramáticas regulares e livres de contexto baseados em modelos adaptativos”, *Master’s Thesis*, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil.
- [25] Cereda, P. R. M. & José Neto, J. (2015) “A recommendation engine based on adaptive automata”, in *Proceedings of the 17th International Conference on Enterprise Information Systems - Volume 2: ICEIS*, pp. 594–601.

- [26] Cereda, P. R. M. & José Neto, J. (2016) “AA4J: uma biblioteca para implementação de autômatos adaptativos” in Memórias do X Workshop de Tecnologia Adaptativa – WTA 2016, 2016, pp. 16–26.
- [27] Iwai, M. K. (2000) “Um formalismo gramatical adaptativo para linguagens dependentes de contexto”, PhD Thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil.

AUTHORS

Newton Kiyotaka Miura is a researcher at Olos Tecnologia e Sistemas and a PhD candidate in Computer Engineering at Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da Universidade de São Paulo. He received his Electrical Engineering degree from Escola Politécnica da Universidade de São Paulo (1989) and holds a master's degree in Systems Engineering from the University of Kobe, Japan (1993). His research interests include adaptive technology, adaptive automata, adaptive devices and natural language processing.



João José Neto is an associate professor at Escola Politécnica da Universidade de São Paulo and coordinates the Language and Adaptive Technology Laboratory of Departamento de Engenharia de Computação e Sistemas Digitais. He received his Electrical Engineering degree (1971), master's degree (1975) and PhD in Electrical Engineering (1980) from Escola Politécnica da Universidade de São Paulo. His research interests include adaptive devices, adaptive technology, adaptive automata and applications in adaptive decision making systems, natural language processing, compilers, robotics, computer assisted teaching, intelligent systems modelling, automatic learning processes and adaptive technology inferences.



INTENTIONAL BLANK

STORAGE GROWING FORECAST WITH BACULA BACKUP SOFTWARE CATALOG DATA MINING

Heitor Faria, Rommel Carvalho and Priscila Solis

Applied Computing Post Degree Program,
University of Brasilia (UnB), Brasilia, DF, Brazil

ABSTRACT

Backup software information is a potential source for data mining: not only the unstructured stored data from all other backed-up servers, but also backup jobs metadata, which is stored in a formerly known catalog database. Data mining this database, in special, could be used in order to improve backup quality, automation, reliability, predict bottlenecks, identify risks, failure trends, and provide specific needed report information that could not be fetched from closed format property stock property backup software database. Ignoring this data mining project might be costly, with lots of unnecessary human intervention, uncoordinated work and pitfalls, such as having backup service disruption, because of insufficient planning. The specific goal of this practical paper is using Knowledge Discovery in Database Time Series, Stochastic Models and R scripts in order to predict backup storage data growth. This project could not be done with traditional closed format proprietary solutions, since it is generally impossible to read their database data from third party software because of vendor lock-in deliberate overshadow. Nevertheless, it is very feasible with Bacula: the current third most popular backup software worldwide, and open source. This paper is focused on the backup storage demand prediction problem, using the most popular prediction algorithms. Among them, Holt-Winters Model had the highest success rate for the tested data sets.

KEYWORDS

Backup, Catalog, Data Mining, Forecast, R, Storage, Prediction, ARIMA, Holt-Winters

1. INTRODUCTION

By definition, backup data is only accessed in case of disaster [1], that is supposed to happen rarely. The first data scientist instinct would be to use this information also as a source for analytic engines, instead of fetching it from original source, without the concurrency with regular corporate workload as suggested by Poelker [2].

However, there is still another backup software information that is overlooked by the authors that has a lot of potential and is the scope of this practical work: the catalog database. It contains, for instance, the file locations from every backed-up platform, the duplicated files list, backup and restore jobs history etc.

The catalog learning can be also used to improve backup service itself, in order to identify error trends and capacity problems. A common challenge today, for example, is primary backup data continues to grow, more systems and data are deemed worth protecting, and backup retention is sometimes elongated as addressed by a recent Gartner Consultancy Report[3]. Digital data has snowballed to a level that frequently leads to backup storage capacity depletion, and it's imperative to predict these bottlenecks timely in order to avoid compromising backup data.

The purpose of the present work is to manifest that ARIMA and Holt-Winters forecasting models can provide the backup storage demand prediction, according to the terminated past backup jobs. In Section 2, we present the State-of-the-Art.

In the Section 3, we present the Related Work. The Section 4 shows the Methodology. In Section 5, we present the Results. Finally, the Section 6 draws some conclusions and final remarks. And Section 7, indicates Future Works.

March 01, 2017

2. STATE-OF-THE-ART

Bacula¹ is an open source backup software[4] whose metadata is stored in a database, e.g.: job logs, termination status, list of copied files with paths, storage media association, etc. According to Sibbald [5], it was the first published backup software to use a Structured Query Language and supports both MySQL² and PostgreSQL³ open database services.

The Database Tables section of the Community Version manual [6] provides its full definition (table names, data types etc.), which is going to be the main researched data set of this work, only possible because of its open format.

According to Box et al. [7], a model that describes the probability structure of a sequence of observations is called a stochastic process that is a time series of successive observations is used in order to forecast the probably of distributions of future ones.

Time series are data series listed in time order [8], usually spaced with the same time frame and represented graphically through line charts. They can be also understood as streaming data with discrete values, and they have many tradition applications: mathematical finance [9], weather forecasting [10], intelligent transport [11] and econometrics [12]. Modernly, the DataMarket project [13] hosts varied time series data such as Crime, Ecology and Agriculture.

Box et al. [7] still describes Autoregressive Integrated Moving-Average (ARIMA) as a process more suitable to non-stationary time series observations (v.g.: stock prices) instead of autoregressive (AR), moving average (MA) and mixed autoregressivemoving average (ARMA). Sato (2013), Pati and Shukla (2014), Wang et al. (2015), wrote recent papers using ARIMA, appearing as a relevant forecasting technique.

¹ <http://blog.bacula.org/>

² <https://www.mysql.com>

³ <https://www.postgresql.org/>

Conforming to Goodwin and others [17], Holt-Winters is an exponential based method developed by C.C. Holt (1957) and P. Winters in (1960). As reported by Kalekar [18], it is used when the data exhibits both trend and seasonality (which are elements likely existent in this project observations). In line with Rodriguez et al. [19], exponential smoothing methods are based on the weighted averages of past observations, with the weights decaying exponentially as the observations get older. Puthran et al. (2014), Dantas et al. (2017), Athanasopoulos et al. (2017) and many other modern forecasting projects rely on Holt-Winters technique.

3. RELATED WORK

Until 2007, relevant backup book authors such as B Little and A. Chapa (2003) and Preston (2007) did not address the possibility of doing data mining in their studied backup software. Probably, they were moved because of the fact their studies mainly focused in proprietary backup software, where their databases have an unknown and closed format that is impossible to be read with third party software.

Guise (2008) was probably the first to write that backup software not only should, but must allow data mining of its metadata (among others): “without these features, a backup product is stagnant and not able to grow within an environment”. For the author, backup software should constitute value frameworks, but never monoliths.

Still, the impossibility of any developer to predict every possible data combination for backup report that a given company needs is also highlighted by Guise: *...the more useful feature of backup software for long-term integration into an enterprise environment is the ability to perform “data mining” - i.e., retrieving from the backup software and its database(s) details of the backups that have been performed.* More recently and even devoid of any scientific method, Poelker (2013) addressed the problem once more, suggesting and enumerating examples of how different types of backup data could be used for Knowledge Discovery and Data Mining, in order to aggregate value to that service.

That said, this work seems to be the very first in order to build data mining models for backup software databases and hopefully the foundation to other further analysis.

4. METHODOLOGY

According to Carvalho et al. (2014), CRISP-DM stands for Cross Industry Standard Process for Data Mining, which consists of a consortium oriented to establish an independent tool and industrial data mining process model. It aims to be generic enough to be deployed in different industry sectors, and sufficiently detailed to contemplate the conceptual phase of data mining project until the results delivery and validation.

Still, according to Carvalho et al., one of the methodology goals is to make data mining projects of any size run more efficiently: in a smaller time frame, more manageable, more inexpensive but the most fruitful.

The proposed life cycle of the data mining is presented in Figure 1 [26]. Data mining life cycle would consist of six phases [27], and their flow is not fixed: moving back and forth between different phases is usually required.

There is also an outer circle that represents the perpetual continuity of data mining process, since information and lessons fetch from a given project will very likely be use in further ones. And the inner arrows suggests the most frequent path between phases.

This will be the methodology used in this project, and the results will follow CRISP-DM phases.

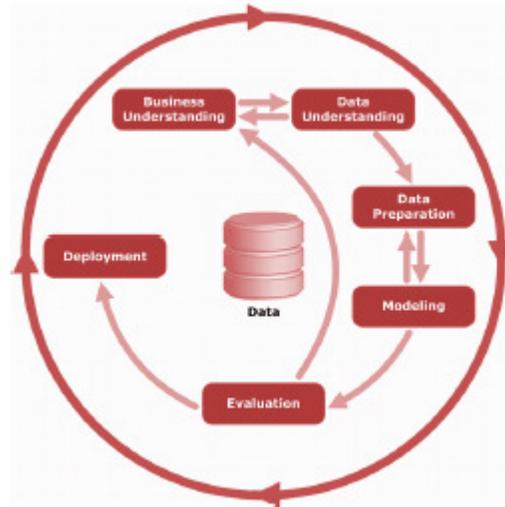


Figure 1. Phases of the CRISP-DM Process Model

There is also an outer circle that represents the perpetual continuity of data mining process, since information and lessons fetch from a given project will very likely be use in further ones. And the inner arrows suggest the most frequent path between phases. This will be the methodology used in this project, and the results will follow CRISP-DM phases.

5. RESULTS

The results are presented ahead as the sections that represent the CRISP-DM executed steps for this work.

5.1. Business Understanding

Acquiring backup demanded storage is not an intuitive task. A naive approach would simply measure storage usage every given time in order to predict its growth. This would ignore already known information about future stored backup behavior, which are the retention times. Retention is the time frame within backup data cannot normally be discarded by the backup software, unless there is an exceptional human intervention. This has a significant impact in storage demand growing prediction, since monthly backup with one year of retention will demand twelve times more data storage occupation than retaining a backup for a single month, for example. A better way to predict backup storage growth is the cumulative sum of all terminated backup jobs during a time frame, subtracted by their own size after their expiration date (when their data is already disposable). For example, if in January 1st a 10GB backup is written, this amount is added to the demanded storage space total. If this job has 1 month retention, the same 10GB value is subtracted in February 1st. The goal is to use already known information to diminish prediction

algorithm margin error, since natural corporate backup size fluctuation demand can be very volatile by itself.

5.2. Data Understanding

Bacula MySQL database, in this case, is accessed using R database interface⁴ and MySQL specific driver⁵. First a test database is used for model developing and testing, then validated with a production database.

Job and Pool tables are exported to a compatible R format. Job table contains the list of terminated backup Jobs information, including their size. Pool⁶ table provides the jobs data retention times.

The job sizes (JobBytes) are expressed in bytes. Null values are discarded and the others converted to Gigabytes, since it is better for human reading. Decimals are rounded with digits, in order to not affect the total cumulative sum.

Backup jobs ending times used to build the time series are expressed originally with date, hours, minutes and seconds (YYYY-MM-DD hh:mm:ss). In order to simplify calculations and because it is insignificant for long term prediction, hour specification was trimmed and only dates are considered.

Retention times in the Pool table (VolRetention) is expressed in seconds. Those are rounded to days, because more significant and usual. Tables are merged so each Job now have their individual retention as a variable value.

Data frame is sort chronologically according to backup termination date, variables that supposed to be known as dates by R are set this way. Job sizes are sum in a cumulative function and a final storage size (sto.size variable) is calculated after the subtraction of already expired backup jobs. Data frame is padded with empty values when there is no backup jobs terminated on those days (necessary for time series). Also, jobs cumulative sum in those days receive last filled row value. There are 95 unique days with backups in this base for further validation reference.

5.3. Modeling

Time series (TS⁷) building R script runs dynamically, fetching first backup week and day from the current data set, in order to be able to work with any other Bacula database and different time frames.

⁴ R DBI Library: <https://cran.r-project.org/web/packages/DBI/>

⁵ RMySQL R Library: <https://cran.r-project.org/web/packages/Rmysql/>

⁶ Pool is the group of backup volumes, or storage units, that has the same attributes such as retention.

⁷ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ts.html>

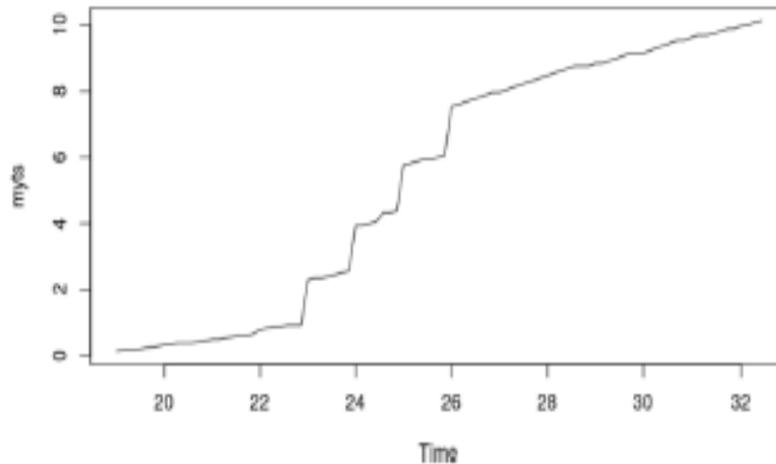


Figure 2. Test database storage size time series: total GB per week.

Figure 2 is a graphical representation of the created time series, and this will be used in order to develop the best models for storage size necessity prediction. It shows a higher initial growth ratio of backup storage size that corresponds to the very beginning of the backup system usage, and happens until backups fill their retention times and start to be recycled (discarded). This will probably be present in lots of Bacula production environments and affects the growing prediction in an undesired way. Since it is a very unpredictable behavior and backup configuration dependent, this is not filtered at this moment. The time scale is expressed in weeks, that is sufficient to run multiple backup jobs but not large enough to ignore huge backup size differences that may happen during a greater period. Last measured storage demand value is 10.1405GB.

5.3.1. ARIMA

In order to build the time series, 180 daily observations were provided, what would give a significant time frame of 6 months of predicted values.

The light gray areas of next plots represents the 80% prediction intervals while dark gray the 95% ones. The blue line is the predicted medium values. Future observation values near the blue line represents higher forecast accuracy.

The forecast library⁸ provides the lower and upper prediction intervals (in this case, with 80% and 95% confidence), which are an estimated range of likely future values.

Figure 3 shows the model seems to be more affected by the initial state backup software operations (first 28 weeks) when the total size of backups grows faster since there are no prior terminated jobs. This can lead to misleading results in fresh production environments and long term predictions. The last forecast value after 6 months of forecast is 25.97GB.

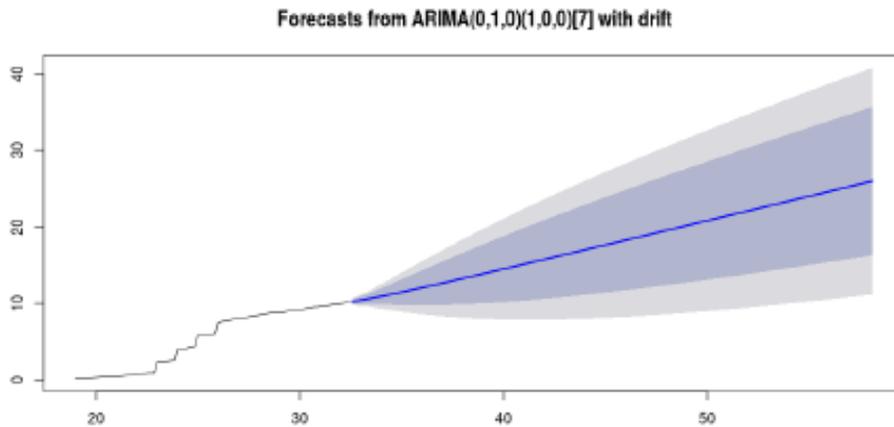


Figure 3. ARIMA model (total GB per weeks).

5.3.1. Holt-Winters

Holt-Winters⁹ exponentially weighted moving averages is more tolerant to the misleading first 28 weeks of the initial state of the backup software, which would provide more reliable prediction values as shown in Figure 4. The forecast value for backup storage for a 6 months time frame is 20.77GB

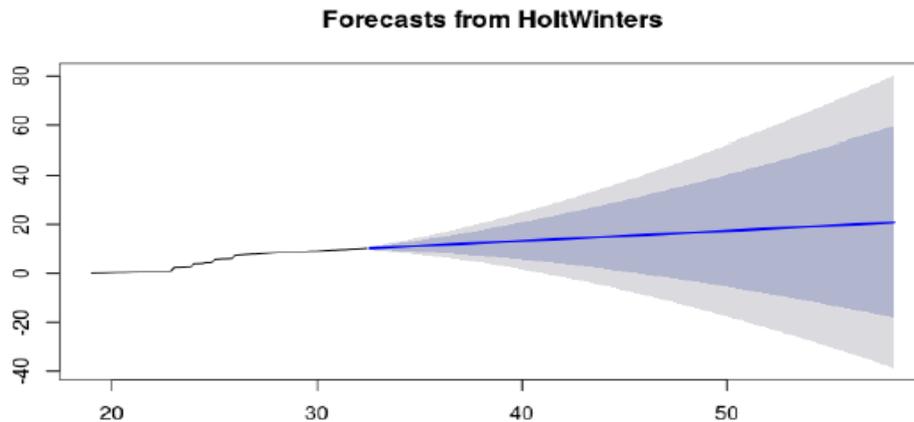


Figure 4. Holtwinters 6 months forecast (total GB per weeks)

5.4. Evaluation

Holt-Winters model responds quicker to trend changes such as lower storage growing trend after initial backup system deploy, being considered more suitable to cope with production environments data. It is the chosen model for testing and validation.

Figure 5 presents, in the same scale, the same used Holt-Winters prediction model against an approximately 30% larger dataset filled with real values. Last forecast storage size is 12.20GB, 10.61% higher than the the actual cumulative backup jobs sum is 11.03GB, but still in the 80% prediction interval.

Another random Bacula production environment database is used in order to apply the selected model. Forecast section corresponds to a slice of approximately 60% entries of the data frame.

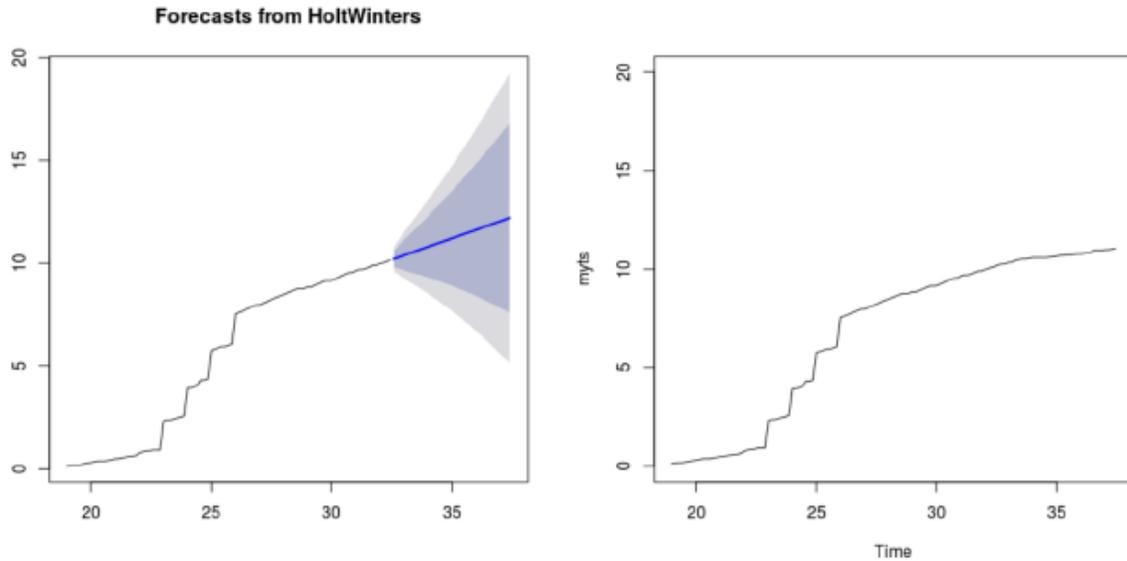


Figure 5. Holt-Winters forecast against test data set storage growing (total GB per weeks).

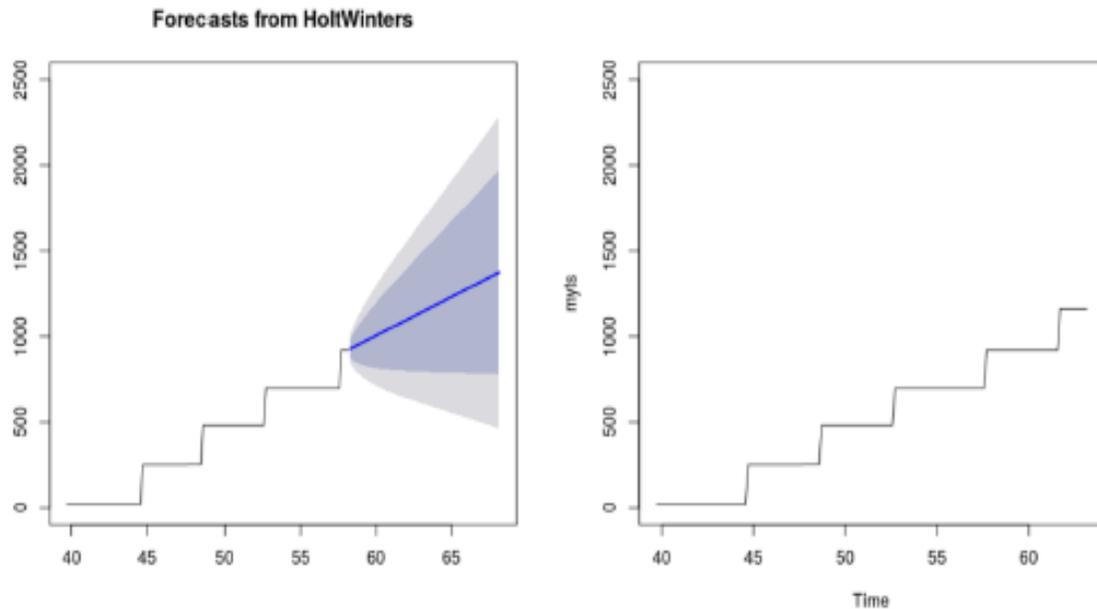


Figure 6. Holt-Winters forecast against production actual data (total GB per weeks).

As displayed in Figure 6, the forecast value (left plot) for ten weeks of prediction (1374.19GB) is 18.33% greater than the actual (right plot) storage size (1161.35GB). However, the lower 95% prediction confidence interval limit for the forecast model is 778.23GB (dark gray area), so the real storage forecast size is into it. The model is satisfactory for available datasets.

6. CONCLUSION

Hyndman [28] study found prediction intervals calculated to include the true results 95% of the time only get it right between 71% and 87% of the time.

In this way, the formula to calculate backup storage cumulative sum for storage size and the choice of the Holt-Winters Model is suitable for the current type of data and for a reasonable and specially vegetative backup size growth, being able to forecast storage growth for a significant amount of time within the 95% prediction interval.

As a remark, it is known that IT infrastructure teams needs to cope with series of unplanned changes from different corporate areas, and for those there are currently no models that could handle them.

The chosen Holt-Winters model must be applied to other production sets of information of different sizes, in order to be considered successfully deployed, which would be the last CRISPDm stage.

Another backup database data mining project execution might also produce the results bellow, among others:

- Predict backup clients demand for restore;
- Fetch all time typical and current successful terminated backup jobs streak;
- Classify backup clients performance (clusterization);
- Identify current backup window compliance;
- Match hardware set suggestions in order to attend current performance and storage demand;
- Analyze the potential benefit of using a file level deduplication feature;
- Analyze the potential benefit of using a block global deduplication feature;
- Identify jobs which behave unexpected with execution log information mining;
- Suggest disk based backup ideal backup volume size.

ACKNOWLEDGEMENTS

Heitor Faria would like to thank Roberto Mourao for the help with the R mechanics and Wanderlei Huttel for providing the production database analyzed in this paper.

REFERENCES

- [1] W. C. Preston, *Backup and Recovery*, 2007.
- [2] C. Poelker, “Backups as a source for data mining,” 2013. [Online]. Available: <http://www.computerworld.com/article/2474646/data-center/backups-as-a-source-for-datamining.html>
- [3] Pushan Rinnen and D. Russel, “Challenging Common Practices for Backup Retention,” Gartner, Inc., USA, Tech. Rep. G00278794, Jul. 2015.
- [4] X. Zhang, Z. Tan, and S. Fan, “NSBS: Design of a Network Storage Backup System,” 2015.
- [5] K. Sibbald, “Bacula Problem Resolution Guide,” Aug. 2015. [Online]. Available: [http://www.bacula.org/7.2.x-manuals/en/problems/Problem Resolution Guide.html](http://www.bacula.org/7.2.x-manuals/en/problems/Problem%20Resolution%20Guide.html)
- [6] —, “Database Tables. Bacula Developer’s Guide,” Aug. 2011. [Online]. Available: [http://www.bacula.org/5.1.x-manuals/en/ developers/developers/Database Tables.html](http://www.bacula.org/5.1.x-manuals/en/developers/developers/Database%20Tables.html)
- [7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Jun. 2015, googleBooks-ID: 1Cy9BgAAQBAJ.
- [8] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms,” in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, ser. DMKD ’03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: <http://doi.acm.org/10.1145/882082.882086>
- [9] M. Corazza and C. Pizzi, “Mathematical and Statistical Methods for Actuarial Sciences and Finance II A skewed GARCH-type model for multivariate financial time series,” 2010. [Online]. Available: <http://booksc.org/book/21982468>
- [10] R. Honda, S. Wang, T. Kikuchi, and O. Konishi, “Mining of Moving Objects from Time-Series Images and its Application to Satellite Weather Imagery,” *Journal of Intelligent Information Systems*, vol. 19, no. 1, pp. 79–93, 2002. [Online]. Available: <http://link.springer.com/article/10.1023/A:1015516504614>
- [11] L. A. James, “Sustained Storage and Transport of Hydraulic Gold Mining Sediment in the Bear River, California,” *Annals of the Association of American Geographers*, vol. 79, no. 4, pp. 570–592, Dec. 1989. [Online]. Available: [http://onlinelibrary.wiley.com/doi/ 10.1111/j.1467-8306.1989.tb00277.x/abstract](http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8306.1989.tb00277.x/abstract)
- [12] J. E. H. Davidson, D. F. Hendry, F. Srba, and S. Yeo, “Econometric Modelling of the Aggregate TimeSeries Relationship Between Consumers’ Expenditure and Income in the United Kingdom,” *The Economic Journal*, vol. 88, no. 352, pp. 661–692, 1978. [Online]. Available: <http://www.jstor.org/stable/2231972>
- [13] “Time Series Data Library - Data provider,” 2017. [Online]. Available: <https://datamarket.com/>
- [14] R. C. Sato, “Gerenciamento de doenas utilizando sries temporais com o modelo ARIMA,” *Einstein (So Paulo)*, 2013. [Online]. Available: <http://www.repositorio.unifesp.br/handle/11600/7670>

- [15] J. Pati and K. K. Shukla, "A comparison of ARIMA, neural network and a hybrid technique for Debian bug number prediction," in *Computer and Communication Technology (ICCCCT), 2014 International Conference on*. IEEE, 2014, pp. 47–53. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/7001468/>
- [16] W.-c. Wang, K.-w. Chau, D.-m. Xu, and X.-Y. Chen, "Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition," *Water Resources Management*, vol. 29, no. 8, pp. 2655–2675, Jun. 2015. [Online]. Available: <http://link.springer.com/10.1007/s11269-015-0962-6>
- [17] P. Goodwin and others, "The holt-winters approach to exponential smoothing: 50 years old and going strong," *Foresight*, vol. 19, pp. 30–33, 2010. [Online]. Available: https://www.researchgate.net/profile/Paul_Goodwin/publication/227439091_The_Holt-Winters_Approach_to_Exponential_Smoothing_50_Years_Old_and_Going_Strong/links/0046351dc5a91a08de000000.pdf
- [18] P. S. Kalekar, "Time series forecasting using holtwinters exponential smoothing," *Kanwal Rekhi School of Information Technology*, vol. 4329008, pp. 1–13, 2004. [Online]. Available: <https://c.forex-tds.com/forum/69/exponentialsmoothing.pdf>
- [19] H. Rodriguez, V. Puig, J. J. Flores, and R. Lopez, "Combined holt-winters and GA trained ANN approach for sensor validation and reconstruction: Application to water demand flowmeters," in *Control and Fault-Tolerant Systems (SysTol), 2016 3rd Conference on*. IEEE, 2016, pp. 202–207. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/7739751/>
- [20] D. Puthran, H. C. Shivaprasad, K. K. Kumar, and M. Manjunath, "Comparing SARIMA and HoltWinters forecasting accuracy with respect to Indian motorcycle industry," *Transactions on Engineering and Sciences*, vol. 2, no. 5, pp. 25–28, 2014. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.437.1043&rep=rep1&type=pdf>
- [21] T. M. Dantas, F. L. Cyrino Oliveira, and H. M. Varela Repolho, "Air transportation demand forecast through Bagging Holt Winters methods," *Journal of Air Transport Management*, vol. 59, pp. 116–123, Mar. 2017. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0969699716302265>
- [22] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos, "Forecasting with temporal hierarchies," 2015. [Online]. Available: <https://mpra.ub.unimuenchen.de/id/eprint/66362>
- [23] D. B Little and D. A. Chapa, *Implementing Backup an Recovery: The Readiness Guide for the Enterprise*, 2003.
- [24] P. d. Guise, *Enterprise Systems Backup and Recovery: A Corporate Insurance Policy*. CRC Press, Oct. 2008, google-Books-ID: 2OtvqySBTu4C.
- [25] R. N. Carvalho, L. J. Sales, H. A. D. Rocha, and G. L. Mendes, "Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil," 2014. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=77038E9B372F7790F8F0FFDE0A3BF3C1?doi=10.1.1.662.1132>
- [26] K. Jensen, "English: A diagram showing the relationship between the different phases of CRISPDM and illustrates the recursive nature of a data mining project." Apr. 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png
- [27] R. Wirth, "CRISP-DM: Towards a standard process model for data mining," 2000, pp. 29–39.

- [28] R. Hyndman, "A state space framework for automatic forecasting using exponential smoothing methods," Jul. 2002. [Online]. Available: <http://robjhyndman.com/papers/hksg/>

AUTHORS

Heitor Faria was entitled with the Alien of extraordinary ability visa by the US Government for his work in Information Technology and Open Source Software. Master degree at Brasília National University (UNB). "Bacula: Open Source Backup Software" (English & Portuguese) and "Open Licenses & Fundamental Rights" books author (Portuguese). Bacula, Alfresco and Wordpress Training instructor at Udemey, with more than 800 students in 46 countries. Works as a System Analyst on a brazilian governmental company called SERPRO and for Neocode Software (Canada). Law Graduated. IT Service Manager and Project Management extension degrees. Bacula brazilian community founder. Has plenty of experience as server/backup systems administrator (Windows, Linux, Netware, directory services) and as IT / Project manager. Speaker at several international open source software events. ITIL-F, TOEFL and LPIC-III certificated professional.



Rommel Novaes Carvalho is a researcher with the Brazilian Office of the Comptroller General and an affiliate professor at the University of Brasília. His research focus is on uncertainty in the Semantic Web using Bayesian Inference, Data Science, Software Engineering, as well as Java, R, and Python programming. He is the developer of PR-OWL (v2.0) and UnBBayes, an open source, java-based graphical editor for Multi-Entity Bayesian Network and Probabilistic Web Ontology Language (PROWL), among other probabilistic graphical models.



Priscila Solis Barreto is professor at the Computer Science Department, University of Brasília. From May/2002 until March/2007, she was a doctoral candidate at the University of Brasília, Department of Electrical Engineering. Her doctoral thesis subject was traffic characterization in multimedia networks. In 2000, she finished her master studies at the Federal University of Goiás, School of Electrical Engineering. Her master thesis subject was in traffic prediction for ATM networks. Priscila worked for several years as a Computer Science Technician and Consultant, developing projects and software for different entities, such as Terra Networks, DR Sistemas and CIAT-BID. She was also a professor at the Catholic University of Goiás for several years. After her undergraduate studies, Priscila had a scholarship to work as a researcher, for one year, in the DIST (Dipartimento de Informática, Sistemística e Telemática), at the University of Genova, Italy.



Her undergraduate studies were made at the Faculty of Engineering in Computer Science, University Francisco Marroquin, in Guatemala City, Guatemala. Currently, Priscila continues her research at the COMNET Lab within the LARF Research Group. Her main research interests are multimedia traffic models, network performance, network planning and traffic engineering. Priscila is also interested in the development of software tools for simulation. Another recent topic she'd began to work is Future Internet and Trust and Security in Wireless Networks.

AUTHOR INDEX

- Abdullah A. Al-Shaher* 39
Ahmad Habboush 79
Alessandro Niccolai 49
Andraws Swidan 105
Ayman A Alhelbawy 01
Edwin R. Hancock 39
Elena Mugellini 113
Gabriel Terejanu 01
George Kour 127
Haim Levkowitz 87
Haroldo Issao Guibu 17
Hassan Noureddine 113
Heitor Faria 185
Hussein Charara 113
Hussein Hazimeh 113
Jawad Makki 113
João José Neto 17, 173
Jörg Hering 147
Kareem E Abdelfatah 01
Kazuo Kiguchi 167
Kazuo Kiguchi 71
Koki Honda 167
Luiz Fernando Capretz 141
Mansour Bader 105
Mariam Abdullah 113
Martin Ciupa 59
Mazin Al-hadidi 105
Mehrdad Nourai 87
Meriam A. Bibile 29
Mohammad M Alnabhan 79
Mohammed Salem Atoum 79
Nemai C. Karmakar 29
Newton Kiyotaka Miura 173
Omar Abou Khaled 113
Pradeep Waychal 141
Priscila Solis 185
Ramez Alkhatib 09
Raz Regev 127
Riccardo E. Zich 49
Rommel Carvalho 185
Shaul Strachan 127
Sho Yabunaka 71
Stefano Frassinelli 49
Takuto Fujita 71
Toshihiro Hara 71
Yusaku Takeda 71