
# ICCGI 2017

The Twelfth International Multi-Conference on Computing in the Global Information Technology

ISBN: 978-1-61208-571-5

July 23 - 27, 2017

Nice, France

**ICCGI 2017 Editors**

Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany

Constantin Paleologu, University Politehnica of Bucharest, Romania

Bobby Law, Glasgow Caledonian University, Scotland

# ICCGI 2017

# Foreword

The Twelfth International Multi-Conference on Computing in the Global Information Technology (ICCGI 2017), held between July 23 - 27, 2017 - Nice, France, continued a series of international events covering a large spectrum of topics related to global knowledge concerning computation, technologies, mechanisms, cognitive patterns, thinking, communications, user-centric approaches, nanotechnologies, and advanced networking and systems. The conference topics focus on challenging aspects in the next generation of information technology and communications related to the computing paradigms (mobile computing, database computing, GRID computing, multi-agent computing, autonomic computing, evolutionary computation) and communication and networking and telecommunications technologies (mobility, networking, bio-technologies, autonomous systems, image processing, Internet and web technologies), towards secure, self-defendable, autonomous, privacy-safe, and context-aware scalable systems.

This conference intended to expose the scientists to the latest developments covering a variety of complementary topics, aiming to enhance one's understanding of the overall picture of computing in the global information technology.

The integration and adoption of IPv6, also known as the Next Generation of the Internet Protocol, is happening throughout the World at this very moment. To maintain global competitiveness, governments are mandating, encouraging or actively supporting the adoption of IPv6 to prepare their respective economies for the future communication infrastructures. Business organizations are increasingly mindful of the IPv4 address space depletion and see within IPv6 a way to solve pressing technical problems while IPv6 technology continues to evolve beyond IPv4 capabilities. Communications equipment manufacturers and applications developers are actively integrating IPv6 in their products based on market demands.

IPv6 continues to represent a fertile area of technology innovation and investigation. IPv6 is opening the way to new successful research projects. Leading edge Internet Service Providers are guiding the way to a new kind of Internet where any-to-any reachability is not a vivid dream but a notion of reality in production IPv6 networks that have been commercially deployed. National Research and Educational Networks together with internationally known hardware vendors, Service Providers and commercial enterprises have generated a great amount of expertise in designing, deploying and operating IPv6 networks and services. This knowledge can be leveraged to accelerate the deployment of the protocol worldwide.

We take here the opportunity to warmly thank all the members of the ICCGI 2017 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICCGI 2017. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICCGI 2017 organizing

committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICCGI 2017 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of computing in the global information technology.

We are convinced that the participants found the event useful and communications very open. We also hope that Nice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

**ICCGI 2017 Chairs:**

**ICCGI Steering Committee**

Carlos Becker Westphall, Universidade Federal de Santa Catarina, Brazil
Constantin Paleologu, University Politehnica of Bucharest, Romania
Elena Ravve, Ort-Braude College - Karmiel, Israel
Ulrich Reimer, University of Applied Sciences St. Gallen, Switzerland
Juho Mäkiö, Hochschule Emden / Leer, Germany
Abdel Lisser, Université Paris Sud, France
Robert Bestak, Czech Technical University in Prague, Czech Republic

**ICCGI Industry/Research Advisory Committee**

John Terzakis, Intel, USA
Isabel Seruca, Portucalense University, Portugal
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Arno Leist, Institute of Natural and Mathematical Sciences | Massey University Auckland, New Zealand
Peter Kieseberg, SBA Research, Austria
Isabel Muench, BSI - German Federal Office for Information Security, Germany
Zeeshan Pervez, University of the West of Scotland, UK

# ICCGI 2017

# Committee

**ICCGI Steering Committee**

Carlos Becker Westphall, Universidade Federal de Santa Catarina, Brazil
Constantin Paleologu, University Politehnica of Bucharest, Romania
Elena Ravve, Ort-Braude College - Karmiel, Israel
Ulrich Reimer, University of Applied Sciences St. Gallen, Switzerland
Juho Mäkiö, Hochschule Emden / Leer, Germany
Abdel Lisser, Université Paris Sud, France
Robert Bestak, Czech Technical University in Prague, Czech Republic

**ICCGI Industry/Research Advisory Committee**

John Terzakis, Intel, USA
Isabel Seruca, Portucalense University, Portugal
Yasushi Kambayashi, Nippon Institute of Technology, Japan
Arno Leist, Institute of Natural and Mathematical Sciences | Massey University Auckland, New Zealand
Peter Kieseberg, SBA Research, Austria
Isabel Muench, BSI - German Federal Office for Information Security, Germany
Zeeshan Pervez, University of the West of Scotland, UK

**ICCGI 2017 Technical Program Committee**

Nadine Akkari, King Abdulaziz University, Jeddah, Saudi Arabia
Konstantin Aksyonov, Ural Federal University, Russia
Cristina Alcaraz, University of Malaga, Spain
Ala Alluhaidan, Claremont Graduate University, USA
Antonova Anna, Ural Federal University, Russia
Michaela Baumann, University of Bayreuth, Germany
Carlos Becker Westphall, Universidade Federal de Santa Catarina, Brazil
Robert Bestak, Czech Technical University in Prague, Czech Republic
Ateet Bhalla, Independent Consultant, India
Dorota Bielinska-Waz, Medical University of Gdansk, Poland
Fernando Bobillo, University of Zaragoza, Spain
Mihai Boicu, George Mason University, USA
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Jean-Louis Boulanger, CERTIFER, France
Xiaoqiang Cai, The Chinese University of Hong Kong (Shenzhen), P. R. China
Maiga Chang, Athabasca University, Canada
Albert M. K. Cheng, University of Houston, USA
Yung Ryn (Elisha) Choe, Sandia National Laboratories, Livermore, USA
Rebeca Cortazar, University of Deusto, Spain

**Copyright Information**

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

**Table of Contents**

# Application of 3D-Dynamic Representation of DNA/RNA Sequences to a Characterization of the Zika Virus Genome

Piotr Wąż

Department of Nuclear Medicine
Medical University of Gdańsk, Poland
Email: phwaz@gumed.edu.pl

Dorota Bielińska-Wąż

Department of Radiological Informatics and Statistics
Medical University of Gdańsk, Poland
Email: djwaz@gumed.edu.pl

***Abstract*—A new method in bioinformatics, which is referred to as *3D-dynamic representation of DNA sequences* is briefly outlined. The aim of this work is an application of this method to the description of the Zika virus genome. We expect to reveal some new features of the sequences comparing to our previous results obtained by using the 2D-dynamic representation.**

***Keywords–Bioinformatics; Alignment-free methods; Descriptors.***

## I. INTRODUCTION

The problem of classification is related to the problem of similarity of the objects. The objects arranged in simple, one-dimensional sets may be classified in a unique way according to a single aspect of similarity. The problem becomes more complicated if we consider multi-dimensional sets, i.e., objects characterized by several different aspects. The degree of similarity (classification to some particular groups) depends on the set of selected aspects, on the number of aspects considered and on the mathematical measure establishing the relations between different properties.

The aim of the studies is a creation of new bioinformatical models carrying information about similarity of the DNA sequences. This information is relevant for solving many biomedical problems. The inspiration for these studies has interdisciplinary character.

A sequence is defined as a sequence of symbols. In the case of the DNA this is a sequence composed of four letters corresponding to four nucleotides: A - adenine, C - cytosine, G - guanine, T - thymine.

The starting point in the methods mentioned above is a DNA sequence represented as a sequence of four symbols in the $5'$ to $3'$ direction. In the methods called in the literature *Graphical representations of DNA sequences*, the sequence of symbols is represented by graphs. The aim of these methods is the creation of both graphs and descriptors representing DNA sequences in a unique way. It may happen that a method can not distinguish between two or more different DNA sequences. In these cases, a nonuniqueness appears, which means that several different sequences are represented by the same graph or by the same descriptor. This kind of nonuniqueness in graphical bioinformatics is called in the English literature *degeneracy*.

Nonuniqueness of the description (degeneracy) is an undesired feature of the method. Removing this feature may be difficult in graphical representation methods. The graphs representing DNA sequences are plotted in two or three-dimensional space. The sequences are long and are composed of four different bases. A reduction of such complicated objects to simple, small, two or three-dimensional graphs corresponding to the perceptional abilities of humans, without a significant loss of information, is difficult and often leads to degeneracy. It is also not obvious how to assign nondegenerate descriptors to such graphs. One of the achievements of the present work is the construction of methods which are either free of this uncertainty or the remaining uncertainty is much lower than in the other formerly known methods.

Recently, we have proposed a new method of comparison of DNA/RNA sequences, called by us *3D-dynamic representation of DNA/RNA sequences* [1] [2]. This method belongs to a group of methods in bioinformatics called *graphical representation methods* (See for reviews [3]–[5]). These methods allow for both graphical and numerical comparison of the considered objects. The sequences are very long, and it is not obvious how to represent them graphically. Each method reveals different aspects of similarity, and therefore new approaches are created.

The aim of this work is an application of this method to the description of the Zika virus genome. We expect to reveal some new features of the sequences comparing to our previous results obtained by using the 2D-dynamic representation.

## II. METHOD AND EXPECTED RESULTS

In *3D-dynamic representation of DNA sequences* method, the sequence is represented by a set of material points in a 3D space [1] [2]. The way of construction of the 3D-dynamic graph is described in [1]. The examples of 3D dynamic graphs representing histone H1 coding sequences of plants and of vertebrates are shown in Figures 1 and 2, respectively. The starting point of 3D-dynamic graphs is the origin of the coordinate system – the coordinates of this points are zeros. Each of the bases is represented by a basis vector. Therefore the dimensions of the graphs and their location contain information about the number of particular bases and about their distribution in DNA sequences.

The name of this method (*3D-dynamic representation of DNA sequences*) is related to the numerical characteristics of the graphs (called in the theory of molecular similarity *descriptors*), which are analogous to the ones used in the dynamics, i.e., coordinates of the centers of mass of the graphs, and moments of inertia of the graphs.

 ISBN: 978-1-61208-571-5

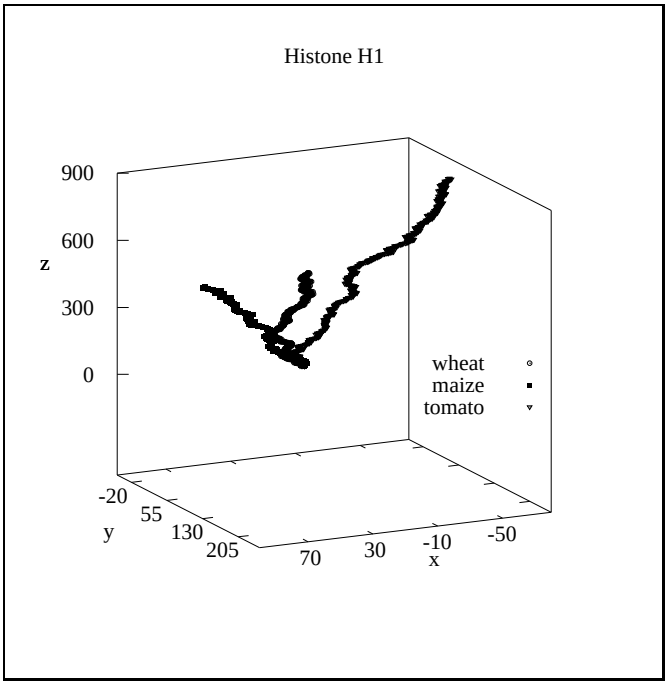



Figure 1. 3D-dynamic graphs.

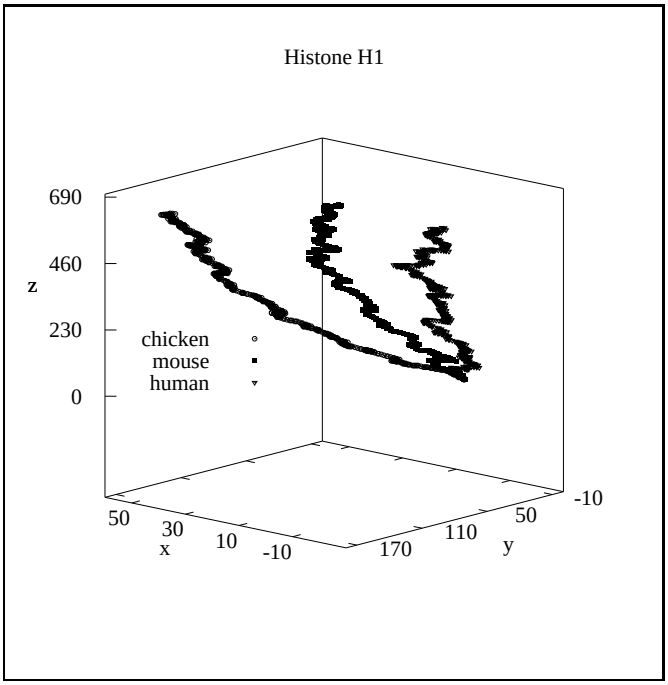



Figure 2. 3D-dynamic graphs.

The coordinates of the center of mass of the 3D-dynamic graph, in the $\{X, Y, Z\}$ coordinate system are defined as [1]

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \quad \mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \quad \mu_z = \frac{\sum_i m_i z_i}{\sum_i m_i}, \quad (1)$$

where $x_i$, $y_i$, $z_i$ are the coordinates of the mass $m_i$. Since $m_i = 1$ for all the points, the total mass of the sequence is $N = \sum_i m_i$, where $N$ is the length of the sequence. Then, the coordinates of the center of mass of the 3D-dynamic graph may be expressed as

$$\mu_x = \frac{1}{N}\sum_i x_i, \quad \mu_y = \frac{1}{N}\sum_i y_i, \quad \mu_z = \frac{1}{N}\sum_i z_i. \quad (2)$$

The tensor of the moment of inertia is given by the matrix

$$\hat{I} = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix}, \quad (3)$$

where the particular matrix elements are defined in [1]. The eigenvalue problem of the tensor of inertia is defined as

$$\hat{I}\omega_k = I_k\omega_k, \quad k = 1, 2, 3, \quad (4)$$

where $I_k$ are the eigenvalues and $\omega_k$ are the eigenvectors. The eigenvalues $I_1$, $I_2$, $I_3$ are called the principal moments of inertia. As the descriptors we select the square roots of the normalized principal moments of inertia:

$$r_1 = \sqrt{\frac{I_1}{N}}, \quad r_2 = \sqrt{\frac{I_2}{N}}, \quad r_3 = \sqrt{\frac{I_3}{N}}. \quad (5)$$

## III. Conclusion and Future Work

In the present work we describe the sequences of the Zika virus genome using 3D-Dynamic Representation of DNA/RNA Sequences. Recently, we have obtained some correlations of the descriptors with time using 2D-dynamic representation of DNA/RNA sequences [6]. Using the present method, some new features of the considered objects are revealed.

Summarizing, using graphical representation methods different aspects of similarity of the DNA sequences, can be considered separately. Only simple objects can be classified in a unique way in terms of their similarity. A pair of complex objects can be similar in one aspect and very different in another one. Using these methods one can indicate properties which are identical or very different for the same pair of the DNA sequences. Graphical representations of DNA sequences constitute both numerical and graphical tools for similarity/dissimilarity analysis of DNA sequences. They can be applied for solving a large class of problems in biology and medical sciences that require such an analysis.

## References

[1] P. Wąż and D. Bielińska-Wąż, "3D-dynamic representation of DNA sequences", J. Mol. Model. vol. 20, 2141, 2014.

[2] P. Wąż and D. Bielińska-Wąż, "Non-standard similarity/dissimilarity analysis of DNA sequences", Genomics vol. 104, pp. 464–471, 2014.

[3] A. Nandy, M. Harle, and S. C. Basak, "Mathematical descriptors of DNA sequences: development and applications", Arkivoc ix, pp. 211–238, 2006.

[4] D. Bielińska-Wąż, "Graphical and numerical representations of DNA sequences: Statistical aspects of similarity", J. Math. Chem. vol. 49, pp. 2345–2407, 2011.

[5] M. Randić, M. Novič, and D. Plavšić, "Milestones in Graphical Bioinformatics", Int. J. Quant. Chem. vol. 113, pp. 2413–2446, 2013.

[6] D. Panas, P. Wąż, D. Bielińska-Wąż A. Nandy, and S.C. Basak, "2D-Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of the Zika Virus Genome", MATCH Commun. Math. Comput. Chem. vol. 77, pp. 321–332, 2017.

 ISBN: 978-1-61208-571-5 

# 2D-Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of Influenza Viruses

Dorota Bielińska-Wąż

Department of Radiological Informatics and Statistics
Medical University of Gdańsk, Poland
Email: djwaz@gumed.edu.pl

Piotr Wąż

Department of Nuclear Medicine
Medical University of Gdańsk, Poland
Email: phwaz@gumed.edu.pl

***Abstract*—The aim of this work is an application of the *2D-dynamic representation of DNA/RNA sequences* for a description of influenza viruses. The descriptors (numerical characteristics) for the considered sequences are calculated. We expect that the results will reveal some new features of the considered objects and will be helpful for the creation of a model of time evolution of influenza viruses.**

***Keywords–Bioinformatics; Alignment-free methods; Descriptors.***

## I. INTRODUCTION

Recently, we have introduced and developed a new method of comparison of Deoxyribonucleic acid/Ribonucleic acid (DNA/RNA) sequences called by us *2D-dynamic representation of DNA/RNA sequences* [1]–[5]. In the 2D-dynamic representation, the DNA/RNA sequence is represented as a set of material points in 2D space ("2D-dynamic graph").The distribution of the points in the plane and the way of calculating their masses is described in [1]. This method belongs to a group of methods in bioinformatics known in the literature as *Graphical Representation Methods*. They allow for both graphical and numerical comparison of the objects. The first methods of this kind have been published in the eighties and nineties [6]–[9]. Since then, many other approaches have been constructed, as for example [10]–[13]. Reviews may be found in [14] [15]. Each method describes different aspects of similarity and still new approaches are constructed.

## II. METHOD AND EXPECTED RESULTS

2D-dynamic representation is based on shifts in a two dimensional space. The DNA/RNA sequence is represented by material points with different masses in a two dimensional space. This method is an improvement of traditional plots, in which particular bases are represented by two orthogonal pairs of colinear basis vectors. Such a choice of the vectors leads to the possibility of shifts back and forth along the same trace. The so called repetitive walks lead to degeneracy: different sequences may be represented by the same graphs. In order to remove the degeneracy, points with masses which are a multiplicity of the unit mass have been introduced. After a unit shift a point with unit mass is localized. If the ends of the vectors during the shifts coincide, then the mass of this point increases accordingly. The total mass of the graph (the sum of all masses) is equal to the length of the sequence.

Several examples of the 2D-dynamic graphs representing different complete genome sequences of Zika virus are shown

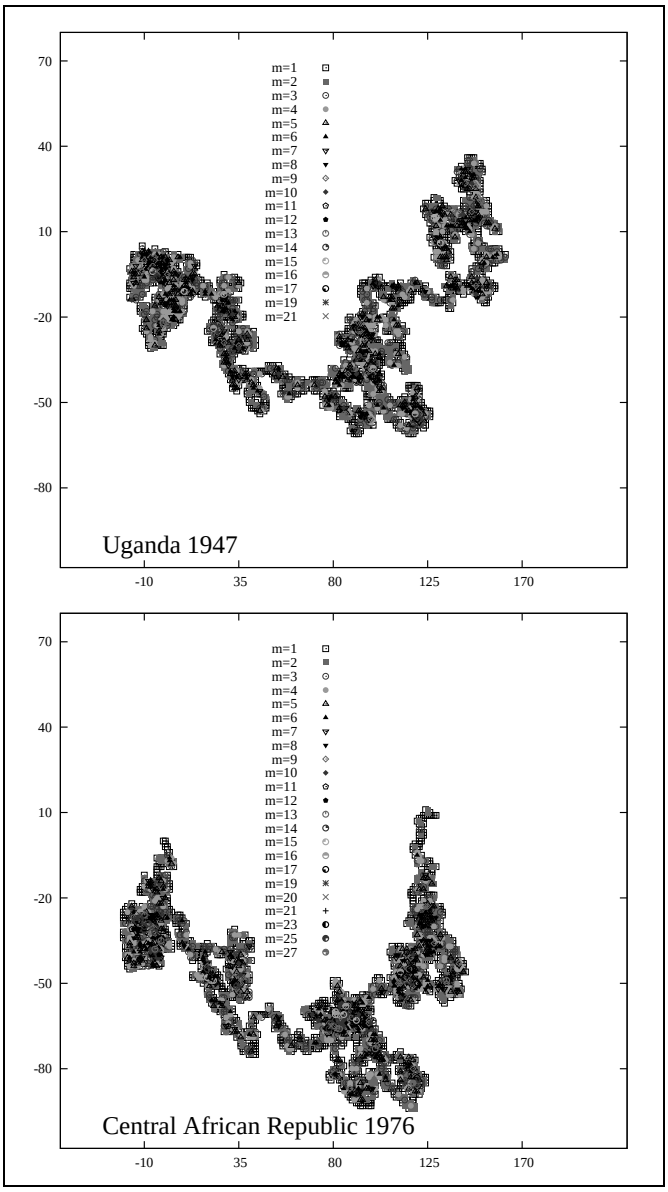



Figure 1. 2D-dynamic graphs.

 ISBN: 978-1-61208-571-5

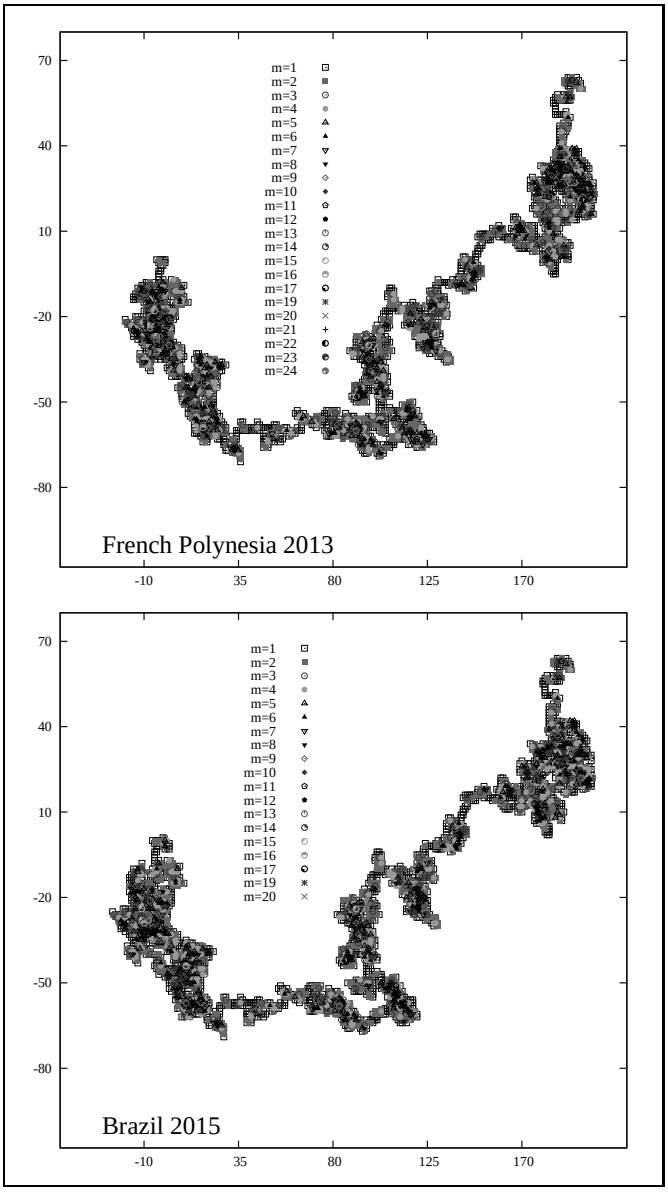



Figure 2. 2D-dynamic graphs.

in Figures 1-2. Similar sequences are represented by similar graphs. The structural forms of the graphs evolve with time. Pairs of graphs are similar to each other: HQ234498 Uganda 1947 is similar to KF268948 Central African Republic 1976 (Fig. 1) and KJ776791 French Polynesia 2013 is similar to KU365777 Brazil 2015 (Fig. 2). This observation is confirmed by the calculations done in [5].

Each graphical object is here described by a set of numerical values called in the theory of molecular similarity *descriptors*. We have shown that dynamical variables of the classical mechanics as, for example, the coordinates of the center of mass are the representative descriptors:

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \quad \mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \tag{1}$$

where $x_i$, $y_i$ are the coordinates of mass $m_i$ assigned to the i-th node of the 2D-dynamic graph. We have also used some other descriptors, as for example the principal moments of inertia of the graphs.

In the present work we apply these descriptors to a description of the sequences of influenza viruses. We hope to discover some correlations between the descriptors and time and place. The necessary data are freely accessible in database Genbank. The mathematical description of these data will be helpful for a general knowledge about the viruses and also for a prediction of their evolution. This information is relevant for the designing of the vaccine.

## REFERENCES

[1] D. Bielińska-Wąż, T. Clark, P. Wąż, W. Nowak, and A. Nandy, "2D-dynamic representation of DNA sequences", Chem. Phys. Lett. vol. 442, pp. 140–144, 2007.

[2] D. Bielińska-Wąż, W. Nowak, P. Wąż, A. Nandy, and T. Clark, "Distribution moments of 2D-graphs as descriptors of DNA sequences", Chem. Phys. Lett. vol. 443, pp. 408–413, 2007.

[3] D. Bielińska-Wąż, P. Wąż, and T. Clark, "Similarity studies of DNA sequences using genetic methods", Chem. Phys. Lett. vol. 445, pp. 68–73, 2007.

[4] P. Wąż, D. Bielińska-Wąż, and A. Nandy, "Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences", J. Math. Chem. vol. 52, pp. 132–140, 2013.

[5] D. Panas, P. Wąż, D. Bielińska-Wąż, A. Nandy, and S.C Basak, "2D-Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of the Zika Virus Genome", MATCH Commun. Math. Comput. Chem. vol. 77, pp. 321–332, 2017.

[6] E. Hamori and J. Ruskin, "H Curves, a novel method of representation of nucleotide series especially suited for long DNA sequences", J. Biol. Chem. vol. 258, pp. 1318–1327, 1983.

[7] M. A. Gates, "Simpler DNA sequence representations", Nature vol. 316, p. 219, 1985.

[8] A. Nandy, "A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes", Current Science vol. 66, pp. 309–314, 1994.

[9] P. M. Leong and S. Morgenthaler, "Random walk and gap plots of DNA sequences", Comput. Appl. Biosci. vol. 11, pp. 503–507, 1995.

[10] Q. Dai, X. Liu, and T. Wang, "A novel graphical representation of DNA sequences and its application", J. Mol. Graph. Model. vol. 25, pp. 340–344, 2006.

[11] H. González-Díaz et al., "Generalized lattice graphs for 2D-visualization of biological information", J. Theor. Biol. vol. 261, pp. 136–147, 2009.

[12] Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, L. Pan, and Y. Ye, "ColorSquare: A colorful square visualization of DNA sequences", MATCH Commun. Math. Comput. Chem. vol. 68, pp. 621–637, 2012.

[13] N. Jafarzadeh and A. Iranmanesh, "C-curve: a novel 3D graphical representation of DNA sequence based on codons", Math Biosci. vol. 241, pp. 217–224, 2013.

[14] D. Bielińska-Wąż, "Graphical and numerical representations of DNA sequences: Statistical aspects of similarity", J. Math. Chem. vol. 49, pp. 2345–2407, 2011.

[15] M. Randić, M. Novič, and D. Plavšić, "Milestones in Graphical Bioinformatics", Int.J.Quant.Chem.vol. 113, pp. 2413–2446, 2013.

 ISBN: 978-1-61208-571-5 

# The EEG Signal Classification in Compressed Sensing Space

Monica Fira
Institute of Computer Science
Romanian Academy
Iasi, Romania
e-mail: mfira@etti.tuiasi.ro

***Abstract*—In this paper, it is analyzed the possibility of the classification of the compressed sensed electroencephalographic (EEG) signals. Compressed sensing is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems. This is based on the principle that, through optimization, the sparsity of a signal can be exploited to recover it from far fewer samples than required by the Shannon-Nyquist sampling theorem. The signals classification is done directly in the compressed space and the EEG signals reconstruction is not necessary. For testing we used EEG signals from a brain computer interface system used for a spelling paradigm. For the classification task, two methods were used, both based on machine learning, namely, Deep learning and Gradient boosting learning.**

***Keywords- EEG; Compressed sensing; BCI; classification; P300***

## I. INTRODUCTION

Compressed sensing (CS), during recent years, was in focus for various fields of science and technology as applied mathematics, computer science, electrical engineering, and signal processing. The novelty introduced by CS is that, in certain conditions, the traditional limits of sampling theory can be overcome. CS relies on the fundamental fact that various signals can be represented using only several nonzero coefficients in a suitable basis of vectors. This basis is named dictionary. Based on the known dictionary and using only very few measurements, such signals can be reconstructed using nonlinear optimization methods. Compressed sensing method is an example of usage in practice of recent mathematical results [1] – [4].

The literature from recent years comprises an impressive number of papers in the field of CS, including 1D and 2D medical signals. Among 1D signals, the most frequently analyzed in connection with CS applications are ECG and EEG since they are most used in the medical world. In the case of EEG signals, there is often a need of records for longer periods of time (i.e., during the night) or for a large number of channels. On the other hand, during the past few years, the human-computer interaction has been thoroughly investigated by researchers from the fields of neurology, psychology and information technology [1] – [8].

Over the past decades, the development of the technology of brain–computer interface (BCI) has provided a novel and promising communication channel for patients suffering from severe motor disabilities, but, being cognitively intact, they need an alternative method to interact with the environment.

As a non-muscular communication and control system, BCI has shown emerging possibilities for people with severe motor disabilities by allowing them to write sentences, move a cursor on the computer screen, play an electronic ping pong game, control an orthosis that provides hand grasp, or operate a brain actuated wheelchair. During the last two decades, BCI electroencephalographic (EEG) based systems have used a variety of electrophysiological signal components: visual evoked potentials, slow cortical potentials, P300 evoked potentials, mu and beta rhythms, and cortical neuronal action potentials [9] [10].

The P300 is a characteristic waveform in the human EEG, occurring as a response to rare task-relevant stimuli in a series of task-irrelevant stimuli. The classical oddball paradigm is usually used to evoke the P300: two categories of stimuli are presented to a subject in random order, one of the categories occurs only rarely and subjects are instructed to determine to which category a stimulus belongs [9] – [11].

The main goal of this paper is to test the following hypothesis: if some data can be classified in the original space, they can be acquired using the CS principle and then they can be classified with approximately similar results in the compressed space. In other words, the close neighbors remain close and far neighbors remain far in compressed space. In other words, the proportion of distances between neighbors is preserved.

The layout of the rest of the paper is as follows: In Section II, there are described the principles of compressed and the mathematical formalism of this method. In Section III is presented P300 spelling paradigm of brain computer interface. In Section IV is described the experimental paradigm used for this work, the subjects and the preprocessing of the data. In Section V, the boosting algorithm and deep learning used for classification are described. The experimental results and conclusions are presented in Section VI.

 ISBN: 978-1-61208-571-5 

## II. COMPRESSED SENSING

Compressed sensing is a rather new paradigm in signal processing that speculates the fact that the so-called sparse signals can be reconstructed from a sma.ll number of projections on a set of random signals. CS applications in biosignals acquisition, compression and processing have been intensely investigated in the last decade [1] – [7].

It is well-known that data acquisition is fundamentally governed by the familiar sampling theorem [8] that states that an $f_0$-bandlimited signal can be recovered from its samples if the sampling frequency is at least 2f0, i.e., twice the highest frequency of the signal spectrum. Thus, in a time window W, an $f_0$-bandlimited analog signal can be represented by $N=2f_0W$ samples equally spaced at $T=1/2f_0$, i.e., as a vector belonging to the space $R^N$. Such a signal can be alternatively described using any complete set of orthogonal functions in $R^N$. Let us observe that sampling is equivalent to taking projections (scalar products) on the elements of the *canonical* basis. In the general case, the signal can be reconstructed from its projections on N orthogonal (or only linear independent) elements in $R^N$ the canonical basis being the most natural particular case and usually the most convenient. Indeed, the above considerations are rigorously valid in the tacit hypothesis of an infinite precision sampling.

On the other hand, there are many (classes of) signals that allow reconstruction based on fewer samples or projections than those required by the sampling theorem. The explanation is that in such cases the samples contain redundant information so that the signals can be compressed and can be reconstructed using projections and *known prior information*. Such a class is that of sparse signals, the *prior* information about them being the fact they admit a representation based on a small number of elements/atoms in $R^N$. A signal is called k-*sparse* if it is known that it can be represented using a number k of elements of $R^N$, the most interesting case being that when k<<N.

Formally, a discrete signal/vector $x \in R^N$ is said to be k-sparse if there exists a basis $\Psi = \{\Psi_i, i = 1,..., N\}$ in $R^N$ such that most of the elements $\alpha = \{\alpha_i, i = 1,..., N\}$ of its representation in that basis, $x = \Psi\alpha$, are zero or, in a more relaxed hypothesis, *approximately zero* so that the signal can be represented *well enough* with the k's largest terms $\alpha_i$ from its expansion with respect to the above basis. In other words, x is said to be compressible, since it can be represented only by the nonzero/largest elements $\alpha_i$. CS theory shows that a k-sparse signal, i.e., which is compressible in a base (or, more general, dictionary) $\Psi$ can be recovered with very good quality from a number *m* of the order of magnitude

$$m = O(k \log(N / k))$$

of non-adaptive linear projections on a set of vectors $\Phi$ which are not coherent with the first, i.e., their elements cannot be used for a compressed representation of any $\Psi_i, i = 1,..., N$.

Thus, instead of measuring the N components of the signal in the canonic base, a number of m $(k < m << N)$ linear projections on the elements of the matrix $\Phi^{N*m}$ are acquired for obtaining the measurement signal (show in Figure 1)

$$y = \Phi x = \Phi\Psi\alpha = \Theta\alpha \qquad (1)$$

where we have neglected the measurement noise. The vectors on which x is projected onto are arranged as the rows of a mxN projection matrix $\Phi$, m < N, where N is the size of x and m is the number of measurements.

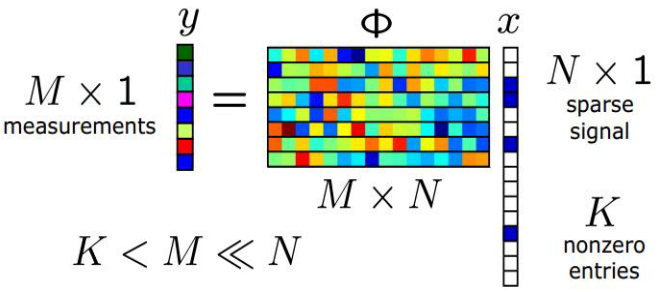



Figure 1. Matrix operations in Compressed Sensing

The system of equations (1) is obviously undetermined since $m < N$, the reconstruction of the initial signal can be made only based on the hypothesis that it is compressible. Under certain assumptions on $\Phi$ and $\Psi$, however, the original expansion vector $\alpha$ can be reconstructed as the unique solution to the optimization problem

$$\hat{\alpha} = \arg\min \|\alpha\|_{l_0} \quad subject\ to \quad y = \Phi\Psi\alpha \qquad (2)$$

the signal is then reconstructed with

$$\hat{x} = \Psi\hat{\alpha} \qquad (3)$$

where $l_0$ is the pseudonorm equal to the number of nonzero elements of $\alpha$, i.e., (2) amounts to finding the sparsest decomposition of the measurement vector y in the dictionary $\Phi\Psi$. Unfortunately, (2) is combinatorial and

 ISBN: 978-1-61208-571-5 

unstable when considering noise or approximately sparse signals. Two directions have emerged to circumvent these problems: (i) pursuit and thresholding algorithms that seek a sub-optimal solution of (2) and (ii) the Basis Pursuit algorithm [1] that relaxes the $l_0$ minimization to $l_1$, solving the convex optimization problem (4) instead of the original one:

$$\hat{\alpha} = \arg_{\alpha} \min \|\alpha\|_{l1} \quad subject\ to \quad y = \Phi\Psi\alpha \quad (4)$$

Many of the results obtained so far in CS refer to "genuine" sparse signals, i.e., to signals that can be represented using precisely k<<N atoms from a given dictionary. However, the results are formally valid for signals that are "approximately sparse", i.e., k is the number of non-negligible elements. Moreover, signals can be sparse in overcomplete dictionaries $\Psi$, i.e., dictionaries with more atoms than their dimension; certain biomedical signals have been found to be sparse in such kind of overcomplete dictionaries. This is the reason why in the past few years, techniques inspired from the mathematic fundamentals of CS have also been applied in the field of biomedical signals, both at the level of processing methods for electroencephalographic (EEG) signals [2] – [4] but also in practical applications [5] including compression, transmission and reconstruction of ECG signals using smart-phones [6].

## III. BRAIN COMPUTER INTERFACE - P300 SPELLER PARADIGM

P300 speller paradigm uses the P300 waves that are expressions of event related potential produced during decision making process. P300 has two subcomponents (as shown in Figure 2 a): the novelty P3 (also named P3a), and the classic P300 (renamed as P3b). P3a is a wave with positive amplitude and peak latency between 250 and 280 ms; the maximum values of the amplitude are recorded for the frontal/central electrodes. P3b has also positive amplitude with a peak around 300 ms; higher values are recorded usually on the parietal areas of the brain. Depending on the task, the latency of the peak could be between 250 and at least 500 ms.

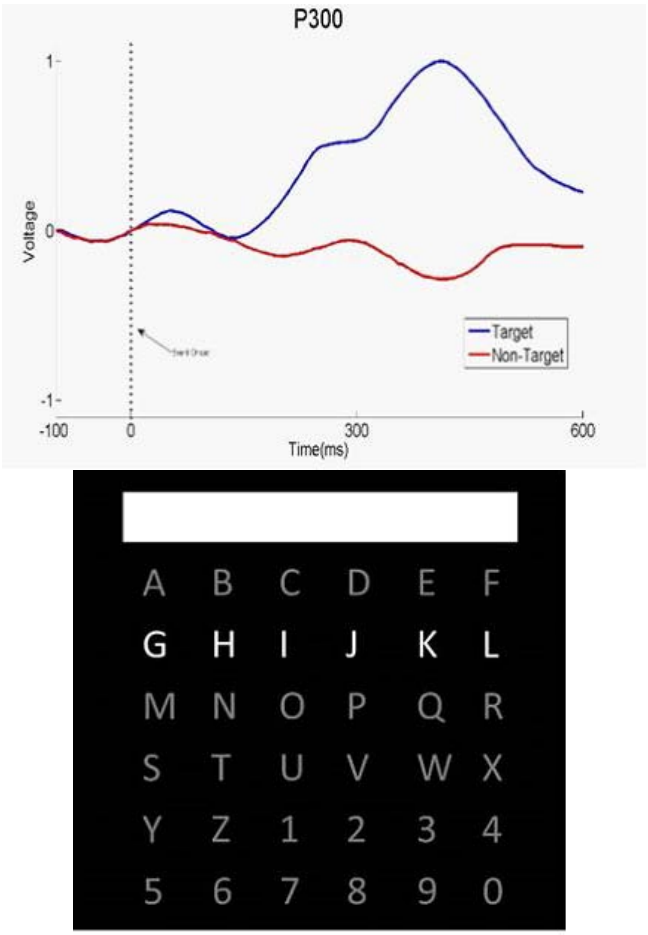



Figure 2. P300 wave and the classical P300 spelling paradigm described by Farwell-Donchin 1988

One of the first examples for BCI is the algorithm proposed by Farwell and Donchin [11] that relies on the unconscious decision making processes expressed via P300 in order to drive a computer. In their approach, a 6x6 matrix (see Figure 2.) of symbols is presented to the user and rows and columns of the matrix are flashed in random order. Subjects can select a symbol from the matrix, by counting the number of times it flashes. Each time the desired character flashes, a P300 is elicited and can be detected by an appropriate algorithm.

## IV. EXPERIMENTAL PARADIGM, DATA ACQUISITION AND PREPROCESSING

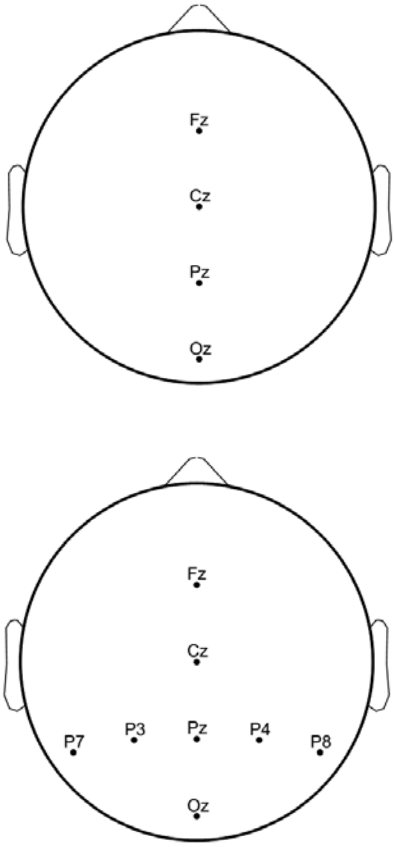
In order to test the classification methods, we used EEG signals collected by Ulrich Hoffmann and collaborators in their laboratory and used by them in the papers [11][15]. The EEG data are available on the internet free at [14].

A setup similar to that described in [11] was used to record and to label the data. A 6×6 matrix containing the letters of the alphabet and the numbers 1-9 was presented to the subjects on a laptop screen. Rows and columns of the matrix were flashed randomly for 100ms with a 100ms

 ISBN: 978-1-61208-571-5 

pause between flashes. Flashes were block randomized, i.e., after 12 flashes each row and column was flashed exactly one time. Two datasets were recorded from each of the subjects on different days. In the first session subjects were asked to spell the french words ”lac,” ”nuage,” ”montagne, ”and ”soleil.” In the second session subjects had to spell the words ”fromage,” ”chocolat,” ”pain,” and ”vin” [12].

Data was recorded from channels FP1, FP2, AF3, AF4, F7, F3, FZ, F4, F8, FC1, FC5, FC6, FC2, T7, C3, CZ, C4, T8, CP1, CP5, CP6, CP2, P7, P3, PZ, P4, P8, PO3, PO4, O1, OZ, O2 with a Biosemi Active 2 system at 2048Hz. The data was then re-referenced to the average of channels O1, OZ, O2, lowpass filtered between 0 and 9 Hz with a 7th order Butterworth filter, and downsampled to 128 Hz. The channels used for re-referencing and channels T7, T8 were not used for further computations because they did not carry relevant information for the detection of P300s. A more detailed description for experimental paradigm, data acquisition, and preprocessing and artifact rejection is presented in [12].



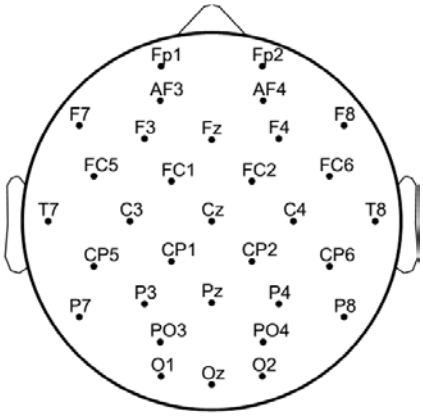




Figure 3. Electrodes configurations

In this paper, it was used a small dataset from this databased with EEG signals.

## V. THE CLASSIFICATION METHODS

In this section the boosting algorithm and deep learning used for classification are described.

### *A. Gradient boosting*

The Gradient boosting classifier from [12] was used. It should be noted that the used software was developed by the authors of that work in order to make a comparison of classification results compared to the original EEG data and the same EEG data but suppose they were purchased used compressed sensed principle.

Gradient boosting is a machine learning method, which builds one strong classifier from many weak classifiers [12].

In [12], Hoffmann and collaborators have described a simple, yet powerful method to detect the P300 from single EEG trials and use it to build a P300 based spelling device. Boosting was employed, to compute from training data a function that detects P300s from single EEG trials. In particular, gradient boosting was used to stepwise maximize the Bernoulli log-likelihood of a logistic regression model. Here ordinary least squares regression was used as weak learner. For a gradient boosting with OLS presented in detail see paper [12].

### *B. Deep learning*

In this paper, we used deep learning in order to learn useful representation of features directly from data. Autoencoder neural networks are able to extract features from unlabeleld data. Autoencoders are used as instruments to train deep neural networks. The training mechanism for autoencoders is considered unsupervised because no labeled data are needed. Autoencoders are trained to replicate their inputs at their outputs by finding a set of weights minimize the corresponding cost function: the error between the inputs and their reconstruction at the outputs [13]. An autoencoder has two parts: an encoder and a decoder. Both, the encoder and decoder could have multiple layers, but, usually, they are designed with a single layer for each of them [13]. The

 ISBN: 978-1-61208-571-5 

training algorithm for autoencoders is back-propagation based.

By cascading two or more autoencoders, a deepnetwork can be obtained.

## VI. Experimental Results and Discussions

The experimental results of classification will be presented in both real acquisition space and in compressed space where the EEG signal will be collected by applying the compressed sensed concepts.

For testing there were used EEG signals collected by Hoffmann, namely, a reduced database available at [13]. This database contains EEG signals collected for 32 channels, grouped in 942 vectors to be classified EEG, lasting 1 sec each.

We have chosen four electrode placement configurations, which we tested for both the original signal and the compressed EEG signals (see Figure 3).

In order to test the classification in the compressed space, we chose two compression ratios, namely, compression of 5: 1 and respectively 10:1. Thus, using compressed sensed algorithm and a random matrix, we simulated that we acquire compressed sensed EEG signals.

For compression evaluation we used the compression rate (CR) defined as the ratio between the number of bits needed to represent the original and the compressed signal.

$$CR = \frac{b_{orig}}{b_{comp}}$$

where $b_{orig}$ and $b_{comp}$ represent the number of bits required for the original and compressed signals, respectively.

### A. Gradient boosting

For testing using the gradient boosting method, the configuration parameters were kept the same as in [12]. Namely, the maximal number of iterations of the boosting algorithm Mmax was set to 200, the optimal M was determined in a 30×10 cross-validation loop, and $\varepsilon$ was set to 0.05.

Table I. The Maximum Classification Rate for Original and Compressed Sensed EEG Signals For Gradient Boosting

| **Gradient boosting method – 23 channels (Fp1, AF3, F3, Fc1, Fc5, C3, CP1, CP5, P3, Pz, PO3, PO4, P4, CP6, CP2, C4, FC6, FC2, F4, AF4, Fp2, Fz, Cz)** | |
|---|---|
| **The classification Space** | **Max Classification rate** |
| EEG originals | 86% |
| CS with 10:1 | 80% |
| CS with 5:1 | 79% |
| **Gradient boosting method – 8 channels (Fz, Cz, Pz, Oz, P7, P3, P4, P8)** | |
| EEG originals | 86% |
| CS with 10:1 | 80% |
| CS with 5:1 | 79% |
| **Gradient boosting method – 4 channels (Fz, Cz, Pz, Oz)** | |
| EEG originals | 81% |
| CS with 10:1 | 75% |
| CS with 5:1 | 73% |

Figure 4 shows the accuracy obtained during the cross-validation loop for configuration with 23 channels. As can be seen, the gradient boosting algorithm converges to an optimal solution. The difference between the rate of classification in the original and compressed sensed space is relatively small, only 6 percent.

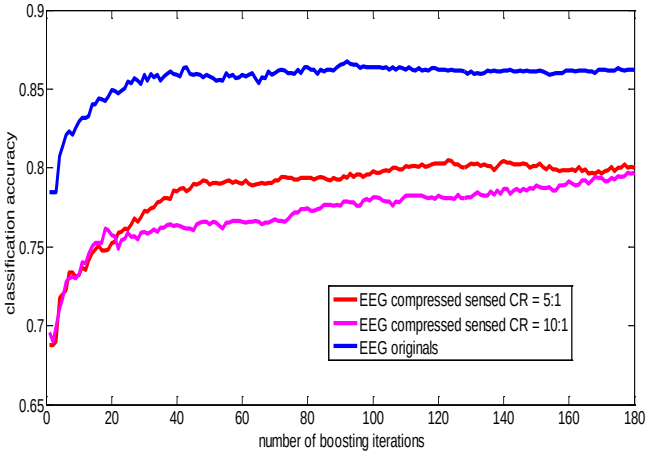



Figure 4. The percentage of classification performance for different values of M.

### B. Deep learning

The relevant features for P300 waves are not directly identifiable on each segment of signal we have available. From this reason we selected as tool for classification the deep networks with auto-encoders which are able to extract relevant features from unlabeled data.

Table II. The Maximum Classification Rate for Original and Compressed Sensed EEG Signals for Depp Learning

| **Deep learning neural network** | | | |
|---|---|---|---|
| **The classification Space** | **Max Classification Rate %** | **Size of signals to be classified** | **Optimum network config.** |
| **Gradient boosting method – 23 channels (Fp1, AF3, F3, Fc1, Fc5, C3, CP1, CP5, P3, Pz, PO3, PO4, P4, CP6, CP2, C4, FC6, FC2, F4, AF4, Fp2, Fz, Cz)** | | | |
| EEG originals | 95% | 2944 | 200-50 |
| CS with 10:1 | 81% | 294 | 50-10 |
| CS with 5:1 | 78% | 588 | 65-10 |
| **Gradient boosting method – 4 channels (Fz, Cz, Pz, Oz)** | | | |
| EEG originals | 80% | 512 | 50-5 |

| | | | |
|---|---|---|---|
| CS with 10:1 | 74% | 52 | 20-5 |
| CS with 5:1 | 72% | 104 | 50-5 |

We used deep networks consisting in two auto-encoders followed by a soft-max layer for the classification of original signals and also for the compressed sensed ones. In case of not compressed signals, the first auto-encoder has 200 hidden elements and the second one has 40. For the compressed signals, we used a first auto-encoder with 40 hidden elements and a second one with only 10.

## VII. CONCLUSIONS

In this paper, it was analyzed the possibility of EEG signals classification (from a spelling paradigm) into EEG signal containing P300 waveform and EEG signals without P300 wave. This classification is the essential element in a BCI system for spelling. Thus, starting from a method proposed by Hoffmann which is based on the gradient boosting classification, it was tested the possibility of the classification of the compressed sensed EEG signals. In other words it was analyzed the possibility of classifying compressed EEG signals, into compressed space which was named compressed sensed space. The utility of this classification is derived from the fact that using the mathematical foundations of CS, the EEG signals can be acquired directly in a compressed form (i.e. the number of samples in the EEG signal is lower than the one indicated by the sampling theorem based on the Nyquist frequency).

It was noticed that using the gradient boosting algorithm, the obtained classification rates have close values for the normal space and the compressed sensed space. Thus, if the original EEG signals classification rate is 86%, for the CS space the classification rate is only 6% lower.

We studied also the classification possibility by using neural network of deep learning type and the results in terms of classification rate are very similar to the gradient boosting method.

The obtained results with both tested methods confirm the hypothesis presented in introduction, according to which the close neighbors in initially space remain close also in the compressed sensed space. This allows the classification of the signals acquired directly in compressed way and it is useful in the applications where only the membership class is important for a signal and not its shape.

## ACKNOWLEDGMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-0832 ***"Medical signal processing methods based on compressed sensing; applications and their implementation."***

## REFERENCES

[1] S. Chen, D. Donoho and M. Saunders, "*Atomic Decomposition by Basis Pursuit*", SIAM Review, 43 (2001), pp. 129-159

[2] L. F. Polania, R. E. Carrillo, Manuel Blanco-Velasco and Kenneth E. Barner, "*ECG compression via matrix completion*", EUSIPCO 2011

[3] L. F. Polania, R. E. Carrillo, Manuel Blanco-Velasco and Kenneth E. Barner, "*Compressed sensing based method for ECG compression*", 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011

[4] C. Wang, J. Liu and J. Sun, „*Compression algorithm for electrocardiograms based on sparse decomposition*", Front. Electr. Electron. Eng. China 2009, 4(1): 10–14

[5] M. Fira and L. Goras, "*Comparison of inter-and intra-subject variability of P300 spelling dictionary in EEG compressed sensing*", International Journal of Advanced Computer Science and Applications , Vol. 7, No. 10, 2016

[6] M. Fira, "*Compressed Sensing of Multi-Channel EEG Signals: quantitative and qualitative evaluation with Speller Paradigm*", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016

[7] N. Cleju, M. Fira, C. Barabasa and L. Goras, „*Robust reconstruction of compressively sensed ECG patterns*", ISSCS 2011 (The 10-th International Symposium on Signals, Circuits and Systems), 30 June – 1 July 2011, Iasi, pp. 507-510, 2011

[8] C. E. Shannon, "*Communication in the presence of noise*", Proc. Institute of Radio Engineers, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[9] J. R.Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller and T. M. Vaughan, "*Brain-computer interfaces for communication and control*", Clin. Neurophysiol. 113 (6), 767-791, 2002

[10] M. A. Lebedev and M. A. Nicolelis, "*Brainmachine interfaces: past, present and future*", Trends in Neurosciences 29 (9), 536 – 546, 2006

[11] L.A Farwell and E Donchin, "*Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials*", Electroencephalography & Clinical Neurophysio,70(6):pp 510-523, 1988

[12] U. Hoffmann, G. Garcia, J.-M. Vesin, K. Diserens, T. Ebrahimi, "*A Boosting Approach to P300 Detection with Application to Brain-Computer Interfaces*", Proceedings of IEEE EMBS Conference on Neural Engineering, 2005

[13] https://www.mathworks.com/help/nnet/ref/trainautoencoder.html#buyr01b-1 (22 May 2017)

[14] http://mmspg.epfl.ch/cms/page-58322.html (22 May 2017)

[15] U. Hoffmann, J. M. Vesin, T. Ebrahimi and K. Diserens, "*An efficient P300-based brain-computer interface for disabled subjects*", J Neurosci Methods. 2008 Jan 15;167(1):115-25. Epub 2007 Mar 13.

 ISBN: 978-1-61208-571-5 

# Terrain Classification Using a Radial Basis Function Network

[1]Tiny du Toit and [2]Hennie Kruger
Computer Science and Information Systems Department
North-West University
Potchefstroom, South Africa
e-mail: [1]Tiny.DuToit@nwu.ac.za, [2]Hennie.Kruger@nwu.ac.za

***Abstract*—In this paper, inertial contact sensor based terrain classification is performed with a Radial basis function network (RBFN). Compared to the more popular Multilayer perceptrons, RBFNs are also intelligent techniques and universal approximators, but with a much simpler structure and shorter training time. It has been shown that RBFNs are efficient classifiers and consequently may be used for terrain classification. For the experiments, a mobile robot platform recorded vibration training data with an inertial measurement unit (IMU) while traversing five different terrains: asphalt, carpet, dirt, paving, and tiles. The composition of these terrains induces specific vibrations in the mobile platform which are measured by the IMU. The vibration signatures are comprised of the mobile robot's linear acceleration, orientation, and the earth's magnetic field. In contrast to most terrain classification techniques found in literature, no pre-processing of the data is performed. This reduces the computational overhead needed for real-time classification. A RBFN is then trained using a hybrid conjugate gradient descent method and *k*-fold cross-validation. Identification of the terrain is performed in real-time. The results are compared to those obtained by a Naïve Bayes method and a Support Vector Machine, which have also been successfully applied to terrain classification in literature. It was found that the RBFN outperformed these other techniques by a relatively large margin. Consequently, the RBFN with no pre-processing of the input data may be used as a contact sensor based terrain classification method.**

***Keywords–classification; IMU; inertial measurement unit; Radial basis function network; sensor; terrain classification.***

## I. INTRODUCTION

Mobile robots are employed in many different operational fields like supply and logistics, surveillance, search and rescue missions, agricultural applications, transportation, cleaning, inspection and entertainment [1][2]. For these operations, it may be necessary to traverse some indoor or off-road terrain, which might influence the vehicle's performance. The efficiency of these vehicles can be improved by detecting their environment. This act of identifying the type of terrain being traversed from a list of candidate terrains such as dirt, sand, or gravel, is called terrain classification [3].

It may be beneficial to identify the current terrain type as the terrain conditions may have an influence on both the motion control and planning stages of the vehicle's trip. Once the mobile robot's control system has knowledge of the surface it is travelling on, it will be easier to maneuver over uneven terrain or around obstacles. In addition, knowledge of the terrain will allow the vehicle to drive at higher speeds. By classifying the terrain, an automated driving process can be obtained which is terrain-dependent.

Research on the identification of terrain types can be divided into two groups: methods relying on noncontact sensors [3] - [6] and methods utilizing contact sensors [7] – [10]. Examples of noncontact sensors are vision sensors and laser scanners. A vision sensor like a charge-coupled device (CCD) camera uses techniques that extract textures and colors from the sensor data to classify it into variable terrains like forests and the sky. Unfortunately, the performance of these techniques is highly dependent on environmental factors like lighting conditions and climate effects and consequently the sensor information can be distorted. Laser scanner sensor data that are obtained from a terrain are converted into frequency information. Learning algorithms then use this information to classify the terrain. A disadvantage of such a method is that it needs numerous data points, which may hinder real-time classification.

Factors like friction, cohesion, damping, stiffness and surface irregularity comprise the terrain interface that is presented to the mobile robot [11]. As the mobile robot traverses the specific terrain, these terrain properties combined with the robot dynamics produce vibrational signatures in body motion. Methods based on contact sensors classify a terrain using sensor information like the vibration frequency or the slope ratio of the mobile robot's body into the terrain type. This enables the mobile robot to choose an appropriate driving mode, which allows the vehicle to traverse the terrain most effectively, prevents physical damage and keep wheels from sinking into the ground.

The goal of this paper is to perform terrain classification using a Radial basis function network (RBFN) as opposed to the well-known Multilayer perceptron (MLP) neural network, which has also been applied to this problem [12]. The MLP that is trained by the backpropagation rule is one of the most used and important neural network models [13]. Owing to its powerful universal approximation capability, the MLP is extensively used in classification, regression, prediction, system identification, control, feature extraction, and associative memory. Broomhead and Lowe [14] proposed the RBFN in 1988. This neural network has become a good alternative to the MLP, since it has equivalent capabilities as the MLP model, but can be trained much faster.

 ISBN: 978-1-61208-571-5 

Previous studies have shown that RBFNs in general are efficient classifiers [1][15]. More specifically, in one study [1] a RBF network has been used for terrain classification where a Discrete Fourier transform was implemented to perform feature extraction. Unfortunately, such pre-processing of the data is a time-consuming task, which may prevent the real-time identification of the terrain. Although the aim of this paper is to investigate the feasibility of a RBFN to perform terrain classification, the results that are obtained will be compared to those achieved by the Naïve Bayes method and the Support Vector Machine (SVM) technique. These two models are also used for terrain classification in literature and the comparison will place the findings in the context of other popular techniques.

Terrain classification will be performed based on real-time vibration data obtained from an inertial measurement unit (IMU) contact sensor. No pre-processing, as reported in previous studies, of the data is performed. The assumption is that the output of the IMU sensor is influenced by the vibrations induced in the platform while traversing different terrains. The test vehicle, a Lego Mindstorms EV 3 mobile robot, is augmented by an IMU mounted on a Raspberry Pi 2 computer. Data that is collected from the IMU on the moving test vehicle is used as the terrain signature. This signature will then be classified by a trained RBFN as one of five predetermined terrains - asphalt, carpet, dirt, paving, or tiles.

The remainder of the paper is organized as follows. In Section II, the relatively simple structure and training of the RBFN will be discussed. A variant of the gradient descent method is used for training. Experiments performed to determine the accuracy of terrain classification using a RBFN will be considered in Section III. The results that were obtained will be examined in Section IV. Finally, some concluding remarks will be presented in Section V.

## II. RADIAL BASIS FUNCTION NETWORKS

In this section, the RBFN architecture and training of the model will be considered.

### A. Architecture

A RBFN is a feedforward neural network with three layers ($J_1 - J_2 - J_3$) [15] – [17] as shown in Figure 1. In the input, hidden and output layers there are $J_1$, $J_2$ and $J_3$ neurons respectively. The bias in the output layer is denoted by $\phi_0(\vec{x}) = 1$ while the nonlinearity at the hidden nodes is denoted by the $\phi_k(\vec{x})$'s. Each hidden layer node uses a Radial basis function (RBF), denoted by $\phi(r)$ for its nonlinear activation function. The hidden layer performs a nonlinear transformation of the input. This nonlinearity is then mapped into a new space by the output layer, which acts as a linear combiner. Normally, all hidden nodes utilize the same RBF; the RBF nodes have the nonlinearity $\phi_k(\vec{x}) = \phi(\vec{x} - \vec{c}_k), k = 1, \dots, J_2$, where $\vec{c}_k$ denotes the center or prototype of the $k$th node and $\phi(\vec{x})$ is an RBF. An extra neuron in the hidden layer can model the biases of the output layer neurons. This neuron has a constant activation function $\phi_0(r) = 1$. The RBFN determines a global optimal solution for the adjustable weights in the minimum mean square error (MSE) sense by using the method of linear optimization. The output of the RBF network, provided by input $\vec{x}$, is given by

$$y_i(\vec{x}) = \sum_{k=1}^{J_2} w_{ki}\phi(\|\vec{x} - \vec{c}_k\|), i = 1, \dots, J_3, \quad (1)$$

where $y_i(\vec{x})$ is the $i$th output, $w_{ki}$ denotes the connection weight from the $k$th hidden neuron to the $i$th output unit, and $\|\cdot\|$ is the Euclidian norm. The RBF usually utilizes the Gaussian function $\phi(\cdot)$ and such a model is normally called the Gaussian RBF network.

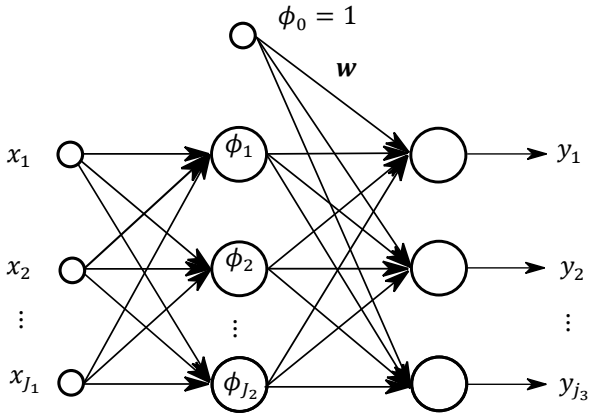



Figure 1. RBF network architecture [16].

Given a set of $N$ pattern pairs $\{(\vec{x}_p, \vec{y}_p)|p = 1, \dots, N\}$, (1) can be expressed in matrix form as

$$\boldsymbol{Y} = W^T\Phi \quad (2)$$

where $\boldsymbol{W} = [w_1, \dots, w_{J_3}]$ is a $J_2 \times J_3$ matrix, $\vec{w}_i = (w_{1i}, \dots, w_{J_2 i})^T$, $\Phi = [\vec{\phi}_1, \dots, \vec{\phi}_N]$ is a $J_2 \times N$ matrix, $\vec{\phi}_p = (\phi_{p,1}, \dots, \phi_{p,J_2})^T$ is the hidden layer output for the $p$th sample, specifically, $\phi_{p,k} = \phi(\|\vec{x}_p - \vec{c}_k\|)$, $\boldsymbol{Y} = [y_1\ y_2\ \dots y_N]$ is a $J_3 \times N$ matrix, and $\vec{y}_p = (y_{p,1}, \dots, y_{p,J_3})^T$.

The RBFN is a universal approximator [16]. If the RBF is appropriately chosen, the RBF network can theoretically approximate any continuous function arbitrarily well. The Gaussian RBF is expressed as $\phi(r) = \exp(-r^2/2\sigma^2)$ where $r > 0$ represents the distance from a data point $\vec{x}$ to a center $\vec{c}$ and $\sigma$ is utilized to control the smoothness of the interpolating function. The Gaussian RBF is a localized RBF with the property that $\phi(r) \to 0$ as $r \to \infty$.

Training of an RBFN is usually performed by a two-phase strategy. During the first phase, suitable centers $\vec{c}_k$ and their corresponding standard deviations, $\sigma_k$, also known as widths or radii are determined. The network weights $\boldsymbol{W}$ are adjusted in the second phase. The training approach that is followed in this research is the supervised learning of all the parameters by the relatively simple gradient descent method.

 ISBN: 978-1-61208-571-5

### B. Training

There is one output unit for each of the five terrain class values (asphalt, carpet, dirt, paving, and tiles). The model trained for the *i*th output unit (class value) is given by:

$$y_i(x_1, x_2, \ldots, x_m) = g\left(w_{i,0} + \sum_{k=1}^{b} w_{i,k} \exp\left(-\sum_{j=1}^{m} \frac{(x_j - c_k)^2}{2\sigma_{global}^2}\right)\right), \quad (3)$$

where the activation function $g(\cdot)$ is a logistic function [18]. A Gaussian RBF network with the same global variance parameter $\sigma_{global}$ for all RBF centers still has universal approximation capability [16]. The appropriate parameter values for $w_{i,k}$ and $\sigma_{global}$ are found by identifying a local minimum of the penalized squared error on the training data. Given $p$ classes, the error function can be expressed as

$$L_{SSE} = \left(\frac{1}{2}\sum_{k=1}^{n}\sum_{i=1}^{p}\left(y_{k,i} - f_i(\vec{x}_k)\right)^2\right) + \left(\lambda \sum_{i=1}^{p}\sum_{k=1}^{b} w_{i,k}^2\right), \quad (4)$$

where $y_{k,i} = 0.99$ if data point $\vec{x}_i$ has the *i*th class value, and $y_{k,i} = 0.01$ otherwise. Instead of using 1.0 and 0.0, the values 0.99 and 0.01 are used to aid the optimization process. Additionally, in (4), $L_{SSE}$, is divided by $n$, the number of training data points, as this was determined through empirical observation to improve convergence with the optimization methods used [19]. Standard calculus is utilized to find the corresponding partial derivatives, which is comprised of the gradients of the error function with respect to the network parameters. Backpropagation is employed to calculate the partial derivatives in the same manner as in Multilayer perceptrons. The hybrid conjugate gradient descent method specified by [20] is used for training.

Before training starts, all numeric inputs in the data are normalized to the [0, 1] interval. This data are transformed back into the original space when predictions are produced. The mode (for nominal attributes) or the mean (for numeric ones) is used to impute missing values. Additionally, nominal attributes are binarized and constant attributes are removed. These same transformations are performed for new inputs when the predictions are made.

Initialization of the network parameters is another important aspect of the training procedure. The initial weights of the output layer are sampled from $\mathcal{N}(0, 0.1)$. This strategy was empirically determined based on the familiar heuristic of choosing small, randomly distributed initial weights [19].

As the *k*-means algorithm is often used to train the hidden layer of the RBFN in an unsupervised process, it is utilized to determine the initial hidden unit centers $c_k$. Furthermore, the initial value of the variance parameter $\sigma_{global}$ is set to the maximum squared Euclidian distance between any pair of cluster centers. This ensures that the initial value of the variance parameter is not too small.

The learning process is accelerated on a multi-core computer by parallelizing the calculation of the error function and its gradient on a user-specified number of threads.

In the next section, the experiments that are performed to determine the RBFN terrain classification accuracy will be discussed.

## III. Experimental Design

The goal of the experiments is to identify the type of terrain being traveled on by a mobile robot, from a list of candidate terrains. Figure 2 shows the Lego Mindstorms EV3 experimental platform used in the investigation. The mobile robot has a Raspberry Pi 2 computer attached to the front with a Sense HAT inertial measurement unit (IMU) in turn connected to the Raspberry Pi. The Sense HAT is readily available and includes the following sensors: A gyroscope, an accelerometer, and a magnetometer. The mobile robot platform is battery powered and moves on rubber treads. An additional battery pack (not shown) is mounted on top and powers the Raspberry Pi computer. The five terrain types used in the study are displayed in Figures 3 to 7.

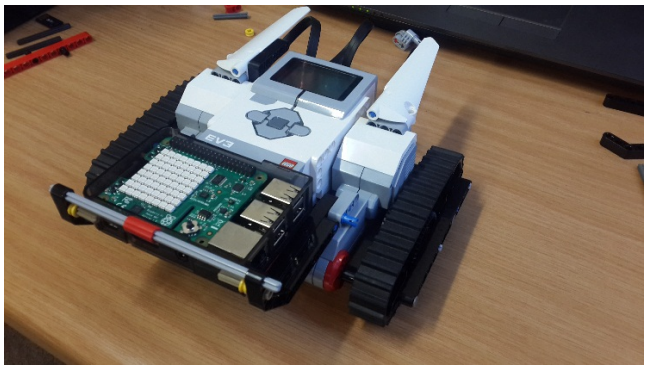


Figure 2. Lego Mindstorms EV 3 mobile robot.

The terrain (asphalt, carpet, dirt, paving, or tiles) on which the mobile robot is currently travelling is identified in real-time. The assumption is that the vibrations induced in the test vehicle and measured by the output of the IMU sensor represent a signature, which can be used to accurately classify the terrains. The data for each terrain is sampled at an irregular rate of $\approx 16\frac{2}{3}$ Hz for a 600 second duration. The RBFN is then trained offline using the RBFN training scheme discussed in Section II (B). Three outdoor terrains (asphalt, dirt, and paving) and two indoor terrains (carpet and tiles) were analyzed.

 ISBN: 978-1-61208-571-5 

Figure 3. Asphalt.

Figure 4. Carpet.

The RBFN architecture for this specific problem has five outputs that serve to identify the terrain type. Each of the output values $y_i \in [0,1]$ denotes the likelihood that a given signal presented as an input to the RBFN matches one of the five candidate terrains. In addition, the RBFN architecture has twelve inputs, which correspond to the dimension of the input signal data point. Each of these input signal data points received from the Sense HAT IMU can be denoted as:

$$[p\ r\ y\ a_x\ a_y\ a_z\ g_x\ g_y\ g_z\ m_x\ m_y\ m_z],$$

where $p, r,$ and $y$ denote the pitch, roll and yaw (measured in degrees), $a$ is the linear acceleration ($m/s^2$) measured in three dimensions ($a_x, a_y$ and $a_z$), $g$ is the rate of turn (degrees/second), also measured in three dimensions ($g_x, g_y$ and $g_z$) and $m$ denotes the earth's magnetic field (gauss), measured in three dimensions ($m_x, m_y$ and $m_z$) of the mobile robot respectively.

Figure 5. Dirt.

Figure 6. Paving.

Figure 7. Tiles.

 ISBN: 978-1-61208-571-5

The Weka system [19] was used for data processing, presentation, classifier training and testing. The terrain classification training dataset contained twelve inputs, five outputs and a total of 49993 samples. For the experiments, 10-fold cross-validation was performed. Results obtained by the RBFN were compared to those found by a SVM model and a Naïve Bayes technique, which are two popular methods found in the literature used for supervised terrain classification [9][10]. In the following section, the results will be discussed.

## IV. DISCUSSION

The classification accuracy results obtained by the experiments are shown in Figure 8.

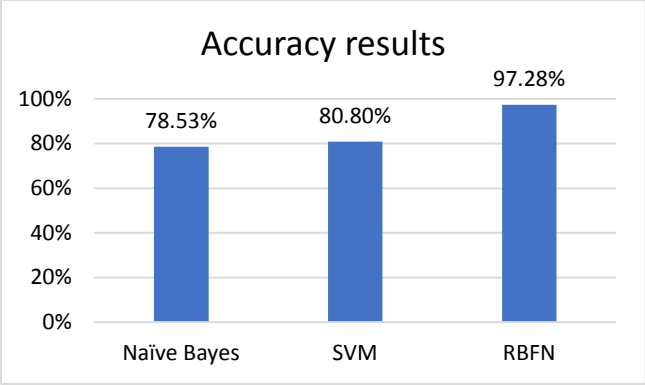



Figure 8. Terrain classification results.

From Figure 8 it can be observed that the machine learning algorithms, ordered from best to worst, are the RBFN, SVM and Naïve Bayes. The latter two techniques produced nearly the same classification accuracy. These results show that the RBFN is a feasible terrain classification technique compared to the other two models and may outperform these techniques by a relatively large margin. This is a promising result as no pre-processing has been performed on the training data.

To summarize, the RBFN applied to terrain classification has the following advantages:

- Compared to the MLP, the RBFN has less model complexity, exhibit better comprehensibility and is easier to construct due to its simpler structure.
- No pre-processing of the input data is performed like in previous studies.
- Classification of the terrain can be performed in real-time because of the onboard IMU contact sensor.
- In terms of predictive accuracy, the RBFN outperformed the Naïve Bayes technique and the SVM model.

Based on these findings, the RBFN is without doubt a technique to consider for terrain classification.

## V. CONCLUSION

In this paper, real-time classification of five given terrains was performed with a RBFN. In contrast to other techniques found in the literature, no pre-processing of the mobile robot platform's IMU vibration sensor data was performed. Eliminating feature extraction reduces the computational overhead needed to identify the terrain in real-time. The results have shown that even without feature extraction, the RBFN is still a feasible model for contact sensor based terrain classification compared to other popular models used for this task. It can be used as an alternative to the MLP model due to its simpler structure and shorter training times. The RBFN has the capability to accurately recognize complex vibration signature patterns and can easily adapt to new terrain signatures by providing the model with new training examples. Unfortunately, compared to the other techniques, offline training of the model can be time consuming.

Future work includes a comparison between the RBFN and MLP models to determine if the RBFN model outperforms the MLP model in terms of terrain classification accuracy. Also, a more detailed comparison with the existing methods must be performed. Metrics like latency (velocity) can be included in the results. Finally, it can be determined if the technique can be applied to other types of robots and how they must be adapted for this task.

## ACKNOWLEDGMENT

The authors would like to thank Mr. Ryno Marx for assembling the mobile robot platform and for acquiring the vibration sensor data for the five terrains.

## REFERENCES

[1] T. Kurban and E. Besdok, "A Comparison of RBF neural network training algorithms for inertial sensor based terrain classification," Sensors, vol. 9, 2009, pp. 6312—6329.

[2] D. Sadhukhan, "Autonomous ground vehicle terrain classification using internal sensors," Florida State University, Master's thesis, 2004.

[3] L. Ojeda, J. Borenstein, G. Witus, and R. Karlsen, "Terrain characterization and classification with a mobile robot," Journal of Field Robotics, vol. 23(2), 2006, pp. 103—122.

[4] A. Angelova, L. Matthies, D. Helmick, and P. Perona, "Fast terrain classification using variable-length representation for autonomous navigation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 2007, pp. 1-8.

[5] A. Talukder et al., "Autonomous terrain characterization and modelling for dynamic control of unmanned vehicles," in Proceedings of the IEEE Conference on Intelligent Robots and Systems (IROS), Lausanne, Switzerland, 2002.

[6] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," Autonomous Robots, vol. 18, 2005, pp. 81–102.

 ISBN: 978-1-61208-571-5 

[7] B. Park, J. Kim, and J. Lee, "Terrain feature extraction and classification for mobile robots utilizing contact sensors on rough terrain," Procedia Engineering, vol. 41, 2012, pp. 846-853.

[8] R. Jitpakdee and T. Maneewarn, "Neural networks terrain classification using inertial measurement unit for an autonomous vehicle," SICE Annual Conference, The University Electro-Communications, Japan, 2008.

[9] C. C. Ward and K. Iagnemma, "Speed-independent vibration-based terrain classification for passenger vehicles," Vehicle System Dynamics, vol. 47, no. 9, 2009, pp. 1095–1113.

[10] M. Happold, M. Ollis, and N. Johnson, "Enhancing supervised terrain classification with predictive unsupervised learning," Robotics: Science and Systems II, University of Pennsylvania Philadelphia, 2006.

[11] F. L. Garcia Bermudez, R. C. Julian, D. W. Haldane, P. Abbeel, and R. S. Fearing, "Performance analysis and terrain classification for a legged robot over rough terrain," "IEEE/RSJ International Conference on Intelligent Robots and Systems", Vilamoura, Algarve, Portugal, October 7-12, 2012.

[12] T. Y. Kim, G. Y. Sung, and J. Lyou, "Robust terrain classification by introducing environmental sensors," IEEE International Workshop on Safety Security and Rescue Robotics (SSRR), 2010.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, D. E. Rumelhart and J. L. McClelland, Eds., vol. 1, pp. 318–362, MIT Press, Cambridge, Mass, USA, 1986.

[14] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," Complex Systems, vol. 2, no. 3, 1988, pp. 321–355.

[15] C. S. K. Dash, A. K. Behera, S. Dehuri, and S.-B. Cho, "Radial basis function neural networks: a topical state-of-the-art survey," Open Computer Science, 6(1), 2016, pp. 33–63.

[16] Y. Wu, H. Wang, B. Zhang, and K.-L. Du. "Using radial basis function networks for function approximation and classification," International Scholarly Research Network, Applied Mathematics, Volume 2012, doi:10.5402/2012/324194.

[17] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, "Neural network design," 2nd edition, Martin Hagan, USA, 2014.

[18] E. Frank, "Fully supervised training of gaussian Radial basis function networks in WEKA," [Online]. http://www.cs.waikato.ac.nz/~ml/publications/2014/rbf_networks_in_weka_description.pdf 2017.03.09.

[19] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA workbench. Online appendix for 'Data mining: Practical machine learning tools and techniques'," Morgan Kaufmann, Fourth Edition, 2016.

[20] Y. H. Dai and Y. Yuan, "An efficient hybrid conjugate gradient method for unconstrained optimization," Annals of Operations Research, 103, 2001, pp. 33-47.

 ISBN: 978-1-61208-571-5 

# Screencasts: Enhancing Coursework Feedback for Game Programming Students

Bobby Law
Dept of Computing, Communications and Interactive Systems
Glasgow Caledonian University
Glasgow, Scotland
Email: robert.law@gcu.ac.uk

***Abstract*—Feedback is an important part of learning and, as such is vital for students to develop and progress throughout their academic life. Programming can be an abstract concept that students find challenging to comprehend therefore good feedback is important to their progress and their motivation to continue programming. This paper will discuss the process of enhancing coursework feedback for Game Programming students through the use of screencasts. The hypothesis being that game programming by its nature is audio-visual thus, providing feedback using an audio-visual medium should increase the students perception of their feedback such that it is perceived to be clearer, easier to comprehend and personalised.**

***Keywords*—Screencasts; Feedback; Software Development.**

## I. INTRODUCTION

The United Kingdom's (UK) National Student Survey (NSS)[1] is a survey for final year students at all of the UK's publicly funded Higher Education Institutions (HEIs) and is administered by Ipsos MORI. The NSS comprises of 27 questions across eight categories attempting to capture the students learning experience. The NSS acts as a barometer of student satisfaction and thus, is an influential survey giving the student body a collective voice. The data from the survey is publicly available and is used by prospective students when choosing their potential University.

This survey has a number of different sections, one of which is Assessment and Feedback. The perennial view from students suggests that there is scope for improvement with regard to Feedback. Comparing all eight categories it can be seen that Assessment and Feedback is continually at the bottom. This would suggest that there is still room for improvement. Table 1 shows all the sections of the questionnaire and their corresponding percentage satisfaction rating. It is noticeable, from Table 1, that satisfaction with Assessment and Feedback is between 5 and 14 percentage points behind 7 of the 8 remaining categories suggesting that the students'impression of feedback and the instrument of feedback delivery have not met entirely with the students'expectations [2][3].

Viewing the statistics on a nation by nation basis against the UK average creates an interesting picture of how students in each of the four nations differ in their perceptions of the level of feedback they receive. Figure 1 shows a comparison of all four nations. Working in an academic institution in Scotland the picture painted is somewhat alarming with Scotland six points below the UK average [4]. The Assessment and Feedback section of the survey is comprised of five questions; two relating to assessment and three relating to feedback. The feedback questions are shown in Table 2. The questions in Table 2 emphasize the students'desire for expeditious, clear and detailed feedback [5].

The remainder of this paper is organized as follows: Section II will provide an overview of the author's rationale for the use of screencasts within the feedback process; indicating the nature of the cohort and the subject area studied. Section III will provide information about pedagogical issues related to screencasting, Section IV offers an introduction to the technologies available for screencasting. Section V presents an overview of the screencasting process, while Section VI reflects on the informally gathered feedback from the student cohort. Section VII discusses issues encountered by the author during the creation of the screencasts. Section VIII attempts to derive conclusions based on the synthesis of Sections III, IV, VI and VII. Section IX offers ideas for future work.

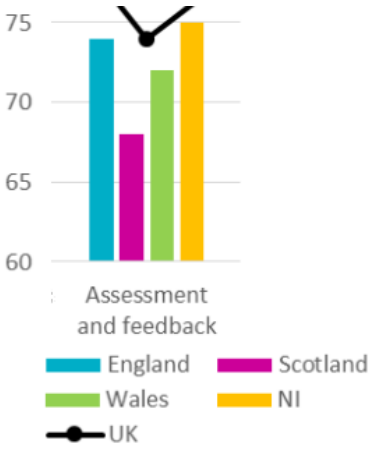



Figure 1. Assessment and Feedback results 2016 by nation

TABLE I. PERCENTAGE SATISFACTION ACROSS CATEGORIES FROM NSS QUESTIONNAIRE

| Categories | 2015 | 2016 |
|---|---|---|
| The teaching on my course | 87 | 87 |
| Assessment and feedback | 73 | 74 |
| Academic support | 82 | 82 |
| Organisation and management | 79 | 79 |
| Learning resources | 85 | 86 |
| Personal development | 83 | 82 |
| Overall satisfaction | 86 | 86 |

TABLE II. EXTRACT OF FEEDBACK QUESTIONS FROM NSS QUESTIONNAIRE

| Feedback Questions asked as part of NSS |
|---|
| Feedback on my work has been prompt. |
| I have received detailed comments on my work. |
| Feedback on my work has helped me clarify things I did not understand. |

 ISBN: 978-1-61208-571-5 

## II. RATIONALE

Teaching programming, and in particular, game programming it can be difficult to offer students feedback on coursework submissions that are not either too generic and brief or ultimately too verbose and overcomplicated. Getting the balance of written feedback correct can be a daunting task. Thompson and Lee [6] suggest that feedback is "a pedagogical tool to improve learning by motivating students to rethink and rework their ideas rather than simply proofread and edit for errors." . Interestingly, Thompson and Lee [6] quote Notar, Wilson and Ross that "feedback should focus on improving the skills needed for the construction of end products more than on the end products themselves". This particular observation is very apt for teaching programming concepts and programming languages as the feedback given is in the context of the students programming skills rather than their end product, in this case their game. The feedback is intended to improve the students ability to produce structured, economical code and illustrate the necessary skills for debugging program code.

The author teaches game programming modules at various levels within the undergraduate programme BSc (Honours) Game Software Development. It would seem natural for game programming students who primarily work with a very audio visual medium to receive feedback for their programming coursework as an audio-visual screencast. It was therefore decided to implement a trial with a second year cohort undertaking the module Game Programming 1. This module was chosen as it was a core module for both the Game Software Development students and the Game Design students. The module introduces students to coding using C++ and OpenGL with the emphasise on the production of a 2D game prototype. The module had approximately 70 students participating in it with a near even split of Game Software Development and Game Design students.

The coursework required the students to create a game of their choosing. The coursework specification provided the students with a number of requirements that had to be met and a marking scheme was provided as a guide to the aesthetic appearance of their game and the functional aspects of the underlying code.

## III. PEDAGOGICAL ISSUES

So what is a screencast? For the purposes of this paper a screencast will be defined as a recording of the current content of the computer screen with an audio narration providing relevant commentary, i.e., feedback [7]–[9]. As part of their learning it is important for students to receive feedback on any of the work that they produce.

Race [10] identifies a number of common formats used to disseminate feedback to students: handwritten, word processed, model answers/solutions, rubric proformas, oral feedback, email and computer marked assessment. These methods can be issued individually or as general feedback based on the performance of a cohort or group.

Race [10] suggests five attributes of feedback: Timely, intimate and individual, empowering, open doors not close them and manageable. Timely feedback is a goal that is highly desired and greatly prized, but, can be dictated by class size or other commitments. Intimate and individual feedback should reflect the student's own submission. Empowering feedback is harder to achieve, as it is a balancing act between positive feedback and a critic, warts and all, of the student's submission. Open doors, not close them refers to the use of language within feedback and the expectation this can set for the student and the feedback they receive for their next submission. Manageable, is viewed from the perspective of both the student and the lecturer, i.e., the effort expended by the lecturer to produce the feedback and the volume of feedback received by the student could cause them to miss something important [10].

Using the written word to provide annotated feedback to students can be taken out of context [9] and therefore the benefit of the feedback can be lost. Worse still, the feedback taken out of context can be misconstrued as a criticism of their work [8] rather than a pointer to improvement. The loss of visual and aural cues, which aid understanding [11], from the written feedback process is therefore something that screencasting can help combat.

As part of Evans [12] "12 pragmatic actions" for effective feedback, one suggestion is for students to be presented with an early assessment opportunity such that they can receive early feedback, which, can be built upon prior to final submission.

It has been mooted that audio-visual screencasts can create for the Lecturer the concept of "social presence" and "an opportunity for conveying positive encouragement through intonation." [8]. This ability to use intonation to emphasize important [2] aspects of feedback make the use of screencasts a benefit for the student. Couple this with the ability to hear the feedback in the manner the Lecturer intended it and the loss of the visual and aural cue associated with face to face feedback are somewhat restored. The volume of information that can be presented to the student via the audio aspect of screencasts is far larger than written feedback alone and in a shorter time period [2][13][14].

Galanos et al. cite the use of screencasts as a method of giving a student personalised feedback by recording the lecturer debugging the students program code while commenting on it [15]. Also suggested is the use of an attached webcam to offer "picture in picture" of the lecturer while debugging the program code, helping to offer that personal touch [15].

It has been suggested that screencasts can aid the student's understanding of their feedback by negating the need for continual cross-referencing between feedback and assessment and secondly the use of conversation style feedback rather than a more formal written academic feedback [8]. It has also been suggested that students find it clearer to "understand the marker's reasoning" [7] and comments [16] when presented in a screencast.

Clarity of feedback is important to students [17]; they do not want to receive feedback that could be deemed "vague, unclear and confusing" [18]. Thus, the audio-visual nature of screencasts can help enrich the feedback pinpointing unambiguously exactly what is being commented on[18]. The promptness or timeliness of feedback is another concern for students as evidenced by the low scores in the National Student Survey[4]. Hope suggests that educators are under an "obligation to provide meaningful feedback within a reasonable timeframe"[2]. Mathisen proffers anecdotal evidence from the field that screencasts can provide more feedback and can be produced in less time[18].

It has been mooted by O'Malley that one of a quartet of criteria needed for feedback to be effective is for it to be personal [19]. Screencasting offers the student personalised feedback that is tailored to their submission [8]. Chewar and

 ISBN: 978-1-61208-571-5 

Matthews state that the use of screencasts to provide feedback allows for more detailed, accurate and robust feedback [20]. Thomson and Lee also suggest that feedback given through the use of screencasts has the capacity to motivate and boost the students engagement with their learning [6].

## IV. TECHNOLOGY

There are a number of different combinations of hardware and software that can be used to create a screencast. The following sections will describe the hardware and software used by the author to create feedback screencasts.

### A. Hardware

To capture good quality audio it is advisable to refrain from using the built-in device microphone but instead opt for a headset or external microphone [2][21][22]. The benefit of using a headset is the consistent distance from the mouth [23] and the ability to position it slightly below the mouth to minimize the noise of breathing [21].

### B. Software

A number of software packages are available and these range from desktop applications to web based applications which, in turn, vary in price from free to hundreds of pounds [23]. Software used for this paper was Screencast-O-Matic a web based application offering a limited version free. The free version allows up to 15 minutes of recording, recording from screen and webcam, the ability to publish to YouTube and the ability to save in popular formats such as .MP4, .AVI and .FLV. It is relatively easy to use [9] and has a very handy countdown before recording begins.

## V. RECORDING SCREENCAST FEEDBACK

Although the screencast in this instance is being created in response to an unknown entity it is still important to apply the rules of creating instructional screencasts by planning [23]. Planning is very important as there will be a number of areas that will require feedback. For the game produced by the students the coursework feedback was broken into the following areas: aesthetics, game play, code structure and compilation. Each of these areas was broken down further with key points: aesthetics covered the games look and feel and interface design; game play covered the ease and enjoyableness of the game, responsiveness of game objects to keyboard/gamepad interaction; code structure covered neatness, use of the fundamental programming building blocks, use of language features, data structures, and the object oriented paradigm; compilation covered the programming compiling and the appropriate use of compilation switches.

Unlike recording a conventional educational screencast there is no need to produce a script [24] as the coursework submissions will not be predictable and a script can depersonalize the feedback and make it feel unnatural [3]. Armed with the marking scheme and the aforementioned plan the process of creating the screencast could be started. A number of considerations were taken into account before commencing the screencast process:

- Determining a location which has a low level of background noise [21] and little chance of being interrupted.
- Use a good quality headset, positioning the microphone slightly below the mouth [21].
- Switch off any software that activates pop ups such as email, Facebook or instant messenger as these could end up being recorded [3].
- Use and stick to the devised plan for consistency.
- Speak naturally and positively [24] making good use of intonation [2].
- Use of the pause button [23] at the end of each section to allow time to gather one's thoughts prior to the start of the next section.

During the recording process all mouse movements and clicks are visible to the viewer as a large coloured circle that will change colour when the mouse button is clicked. This is exceptionally useful for giving the student unambiguous and precise feedback on their user interface design and layout pointing out what is considered good and what needs improving.

The neatness and compactness of the actual code itself is an important aspect of any programming thus, the screencast gave the author the ability to highlight selected code within the Integrated Development Environment (IDE), in this case Microsoft Visual Studio, offering an audio narrative explaining clearly any deficient code and a visualisation of how the code could be reworked in order to make it neater and more efficient. Good examples of student could also be highlighted and the student commended for its use.

For student submissions that did not execute a debug process could be illustrated that would hopefully allow the student to solve a similar problem if encountered again. This ability to show a debug process in operation is a valuable process that merits a role out to all students as the ability to debug code is a valuable skill.

Most of the screencasts were between 5 and 10 minutes in length depending on the game produced and the exhibited programming ability of the student, which is in keeping with the surveyed literature. The feedback screencasts were then subsequently compressed into a .zip file and returned to the student via e-mail.

## VI. STUDENT FEEDBACK

Initial feedback from the students was, on the whole, positive and helpful with regard to refining the screencast feedback process. Comments were elicited from students in an informal manner. Students were asked to write a short paragraph giving their initial impression of receiving feedback in this manner. As all students received both written feedback and feedback in the form of screencasts this allowed the students to compare and contrast the two forms of feedback proffering their thoughts. From the respondents, the overwhelming feeling was the sense of personalization and tailoring of feedback to their needs. Students were also receptive to the visual code analysis they received indicating that they understood more readily the need for well written, neat and compact code. Although, anecdotal, the quotes from students help to articulate their view of screencasts for feedback.

> *"... felt like the feedback was personal to my work."*
>
> *"I could see what Bobby was talking about and this helped me better understand how my code could be improved."*
>
> *"Helped me with debugging especially break points."*

Based on this positive feedback an depth and more rigorous case study will be undertaken in the next academic year to provide a quantitative measure of the worth of screencasting

 ISBN: 978-1-61208-571-5 

as a delivery mechanism for feedback.

## VII. Issues

From the perspective of the lecturer there are some issues that need to be addressed. Firstly, the time taken to prepare the screencast feedback does not necessarily equate to the actual time of the screencast that the student will observe. This is not, necessarily, due to the screencast being edited but the time taken to record the screencast itself. Although, in Section V, a key piece of advise is to plan and prepare for the screencast by using some form of rubric, the application of this rubric can leave the recording having a staccato and unnatural feel. A solution to this is to pause the recording after each section and compose oneself before recording the next section. This will add time to the process but will prove worthwhile in the long term.

Secondly, choosing a suitable location to record the screencasts is imperative as interruptions not only break the lecturers concentration but also can be inadvertently recorded thus, requiring the recording to be edited or, worse still, to be scrapped. A quiet location devoid of interruptions is not always possible in a busy University. It is not an insurmountable challenge but definitely something to be aware of prior to starting any recordings.

A third issue is the size of the recorded screencasts with regard to the required disk storage. The size is dependant on a number of factors including: video codec used, screen size being recorded, and resolution of recording. For example a screencast recorded using the H.264 video codec for YouTube with a definition of 720p, a resolution of 1280x720, 25 frames per second and lasting 5 minutes will require approximately 1.73 gigabytes of disk space. Thus, for a cohort of 70 students, approximately 121 gigabytes of disk storage would be required. This leads to a secondary issue with the delivery mechanism used for distributing the recordings to the students. Distribution by email can be a problem as there may be a restriction on the maximum file size that can be attached to an outgoing email. If this is the case then an alternative method will be required; this could be by uploading the file to a Managed Learning Environment (MLE).

All of the aforementioned issues are solvable with a bit of careful planning and preparation prior to embarking on the recording process.

## VIII. Conclusion

Results from this pilot project suggest that screencasts could be potentially of benefit to both students and staff. If so, this would go along way to addressing the students perception of feedback as highlighted by the UK's National Student Survey.

Reflecting on the creation of the feedback screencasts, it is an interesting exercise to return to the five attributes of feedback, as defined by Race [10], and attempt to analyse, albeit subjectively, if screencast feedback can be thought of as improving the attribute.

Timely feedback can be considered as a property of the turnaround time from student submission of coursework to the lecturer returning feedback to the student; to this end screencasting has no influence on this attribute. Intimate and individual feedback is an interesting attribute; screencasts can help to achieve this attribute, especially for programming, as the student will receive feedback on their programming code, hearing and seeing the lecturer discuss various aspects of their game's code. Empowering feedback is a balance between providing positive feedback and being able to critic the student's work in such a manner that they feel engaged and enthused to progress and push forward. Screencasting feedback can provide the student with the necessary aural and visual cues to afford them the understanding of what is good with their work but also, in a positive manner, how their work can be improved. This is especially good for programming as it is important for students to understand that code that works can still be improved to make it more efficient and that this is a learning process and not a criticism. Open doors, not close them is a delicate area but with a judicious use of appropriate language and the correct vocal intonation the student can be presented with aural cues and, to a certain extent, visual cues that will allow them to synthesise the intended tone of the feedback. Finally, Manageable, as noted by Race[10] has two aspects: the level of work involved for the lecturer and the volume of feedback given to the student. With regard to the level of work involved for the lecturer this may fluctuate depending on the cohort and the quality of their submissions, therefore, it is possible that it could add somewhat to the lecturers overhead for producing feedback. However, for students, they should have a targeted and enhanced quality of feedback which should not overburden them but provide the important aspects of the desired feedback they need to progress and improve.

The increased feedback that can be crammed into a 5 minute screencast is more personal, clearer and less ambiguous than traditional written feedback. The student can play and replay the video as many times as they like and the feedback will always be viewed as it was intended. The time to produce the screencasts varies by student submission but on the whole it was surprisingly quick in comparison to written feedback of the same depth.

## IX. Future Work

The intention is to repeat the screencast feedback in the next academic year. The number of students undertaking the module will, again, be in the region of 60 students and should offer a suitable number for judging the timeliness of producing feedback screencasts. The hypothesis is that the experience from this first large scale implementation will lead to a more effective and quicker production process for each screencast and the students will benefit from clear, concise and helpful feedback. The module is 12 weeks in duration and students will be asked to submit work at the end of week 8 and also at the end of week 12. Screencast feedback on their week 8 submission will be returned by week 10, which, should allow for the students to benefit from the feedback prior to their final submission in week 12 [12]. After receiving the feedback screencasts the students will be surveyed to ascertain a better representation of their feeling towards this feedback mechanism. Screencast feedback will be returned approximately 10 working days after week 12 submission and should serve to inform the students of their programming progress. The intention is to survey the students again at the end of the module in an attempt to better understand their opinion of screencasts as a means of delivering feedback. The survey will attempt to elicit the students perceptions of the screencast feedback based on the categories of engagement,

 ISBN: 978-1-61208-571-5 

quality and quantity of feedback, helpfulness and comparison to written feedback.

## REFERENCES


[1] N. U. of Students (NUS), "The nation student survey," 2015, [retrieved: July, 2017]. [Online]. Available: http://www.thestudentsurvey.com/

[2] S. A. Hope, "Making movies: The next big thing in feedback?" Bioscience Education, vol. 18, no. 1, 2011, pp. 1–14.

[3] K. Haxton and D. McGarvey, "Screencasting as a means of providing timely, general feedback on assessment," New Directions, vol. 7, 2011, pp. 18–21.

[4] N. U. of Students (NUS), "Nss 2015 national headlines," 2015, [retrieved: July, 2017]. [Online]. Available: http://www.thestudentsurvey.com/

[5] R. Law, "Using screencasts to enhance coursework feedback for game programming students," in Proceedings of the 18th ACM conference on Innovation and technology in computer science education. ACM, 2013, pp. 329–329.

[6] R. Thompson and M. J. Lee, "Talking with students through screencasting: Experimentations with video feedback to improve student learning," The Journal of Interactive Technology and Pedagogy, vol. 1, no. 1, 2012.

[7] M. Robinson, B. Loch, and T. Croft, "Student perceptions of screencast feedback on mathematics assessment," International Journal of Research in Undergraduate Mathematics Education, vol. 1, no. 3, 2015, pp. 363–385.

[8] K. Edwards, A.-F. Dujardin, and N. Williams, "Screencast feedback for essays on a distance learning ma in professional communication," Journal of Academic Writing, vol. 2, no. 1, 2012, pp. 95–126.

[9] G. Stieglitz, "Screencasting: Informing students, shaping instruction," UAE Journal of Educational Technology and eLearning, vol. 4, no. 1, 2013, pp. 58–62.

[10] P. Race, "Using feedback to help students to learn," HEA, York, 2001.

[11] K. Mathieson, "Exploring student perceptions of audiovisual feedback via screencasting in online courses," American Journal of Distance Education, vol. 26, no. 3, 2012, pp. 143–156.

[12] C. Evans, "Making sense of assessment feedback in higher education," Review of Educational Research, vol. 83, no. 1, 2013, pp. 70–120. [Online]. Available: http://dx.doi.org/10.3102/0034654312474350

[13] M. Henderson and M. Phillips, "Video-based feedback on student assessment: scarily personal." Australasian Journal of Educational Technology, vol. 31, no. 1, 2015, pp. 51–66.

[14] F. Harper, H. Green, and M. Fernandez-Toro, "Using screencasts in the teaching of modern languages: investigating the use of jing® in feedback on written assignments," The Language Learning Journal, 2015, pp. 1–18.

[15] R. Galanos, W. Brand, S. Sridhara, M. Zamansky, and E. Zayas, "Technology we can't live without!: revisited," in Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education. ACM, 2017, pp. 659–660.

[16] J. West and W. Turner, "Enhancing the assessment experience: improving student perceptions, engagement and understanding using online video feedback," Innovations in Education and Teaching International, 2015, pp. 1–11.

[17] P. Marriott and L. K. Teoh, "Using screencasts to enhance assessment feedback: Students' perceptions and preferences," Accounting Education, vol. 21, 2012, pp. 583–598.

[18] P. Mathisen, "Video feedback in higher education–a contribution to improving the quality of written feedback," Nordic Journal of Digital Literacy, vol. 7, no. 02, 2012, pp. 97–113.

[19] P. OMalley, "Screencasting and a tablet pc–an indispensable technology combination for physical science teaching and feedback in higher and further education," in Aiming for excellence in STEM learning and teaching: Proceedings of the Higher Education Academy's First Annual Learning and Teaching STEM Conference, 2012.

[20] C. Chewar and S. J. Matthews, "Lights, camera, action!: video deliverables for programming projects," Journal of Computing Sciences in Colleges, vol. 31, no. 3, 2016, pp. 8–17.

[21] P. Smith, "Screencasting as a means of enhancing the student learning experience," Learning and Teaching in Action, 2014, p. 59.

[22] D. Wolff-Hilliard and B. Baethe, "Using digital and audio annotations to reinvent critical feedback with online adult students," International Journal for Professional Educators, 2014, p. 40.

[23] S. Mohorovicic, "Creation and use of screencasts in higher education," in MIPRO, 2012 Proceedings of the 35th International Convention. IEEE, 2012, pp. 1293–1298.

[24] L. A. Jones, "Losing the red pen: Video grading feedback in distance and blended learning writing courses," Association Supporting Computer Users in Education Our Second Quarter Century of Resource Sharing, 2014, p. 54.


 ISBN: 978-1-61208-571-5 

# Towards an Analysis and Evaluation Framework for In-Memory-based Use Cases

Stephan Ulbricht*, Marek Opuszko*, Johannes Ruhland*, Martin Thrum*
*Friedrich Schiller University Jena
Department of Business Informations, Jena, Germany
Email: stephan.ulbricht@uni-jena.de, marek.opuszko@uni-jena.de, johannes.ruhland@uni-jena.de, martin.thrum@uni-jena.de

***Abstract*—The aim of this work is to introduce a framework for the analysis and evaluation of potential In-memory applications. As a base for the framework, relevant influencing factors for the use of In-memory systems were identified. For the evaluation and the identification of further influencing factors, an expert survey was carried out. The results show that aspects relevant to companies were not considered in the past. Therefore, a structured analysis framework is introduced considering also economical factors. The use of the framework and the interpretation of the results will be clarified in the end using selected fields of application.**

***Keywords–In-Memory IT-Systems; Big Data; Business Value; Case Study; In-Memory Computing.***

## I. INTRODUCTION

In December 2014, Amazon introduced the ”Prime Now” service, which guarantees the delivery of several thousands of products within an hour [1]. In the field of high frequency trading, fractions of a second can determine profit or loss [2]. Sociologists have been talking about this subject as the ”age of acceleration” for quite some time [3]. Never before in history were decision makers forced to make entrepreneurial decisions under greater time pressure than today. Furthermore, increasingly huge and heterogeneous data sets are challenging companies. One of the most promising technologies for solving these challenges are In-memory-based IT systems (IMIS). Although the technology was subject to high expectations in the past, the predicted boom has not yet begun. In this context, many companies complain about the lack of useful and economical application scenarios [4][5]. In a study by the American SAP user group, this point is mentioned as one of the main causes for the delayed distribution [6]. The reasons for this is, among others, the previous focus on technical aspects [7]. A study by the market research company PAC [8], on the other hand, shows that the In-memory technology is of great interest to many companies and can play an important role in the future. 36% of the surveyed company representatives see this technology as an important building block in future IT landscapes. In this field of tension, it becomes clear that the In-memory technology has great potential that has not yet been exploited.

In this work, we will introduce a design science based system, able to identify and evaluate influential factors for potential application scenarios of IMIS. The aim of this approach is to examine existing as well as potential future scenarios. Based on an analysis framework, the requirements and their feasibility of use cases are examined. In order to identify possible influencing factors of In-memory application scenarios, case studies and scientific literature are analyzed. Subsequently, the influence factors found are evaluated with the help of field experts who participated in an expert survey, also identifying yet unknown and additional factors.

The paper is organized as follows. Section II introduces the technical background, the existing literature about IMIS and the general methodology. Section III presents in the first part the results of the literature review and the expert survey. In the second part of section III the conceptual framework is introduced and its application is shown. The final section summarizes the contributions to practice and research.

## II. RESEARCH BACKGROUND

The idea of using main memory to store data is not new. These concepts were introduced in the 1980s and 1990s [9][10]. At that time, the main focus was very fast response times which were realized by main memory databases. Due to high costs and low memory sizes, the interest regarding In-memory databases decreased and the technology almost fell into oblivion. With the introduction of the HANA platform [11], SAP has once again placed the focus on IMIS. The previous concerns about the durability of the stored data could be eliminated by the use of non-volatile RAM [12]. The concept of IMIS includes more than a pure data storage in the main memory. In contrast to conventional relational databases, the data is no longer stored row-based, but mainly column-based [13]. The advantage of a column-based storage is on the one side a better data compression and on the other side a better suitability for analytical tasks. Originally, the main application area of IMIS were fast and flexible analysis of large amounts of data in data warehouses. In the meantime, the application areas were extended to transaction systems. The goal here is to dissolve the historically grown separation between online analytical processing (OLAP) and online transaction processing (OLTP) systems [14][15]. These hybrid systems are referred to as Online Mixed Workload Processing (OLXP) [16] and Hybrid Transactional/Analytical Processing (HTAP) [17].

The advantage of a common data storage is the elimination of ETL processes from the OLTP into the OLAP system. In addition, transactional data can be used for analytical and planning tasks. Furthermore, there is a potential for savings through the elimination of an additional system [18]. However, it is important to note that analysis and transaction systems have fundamentally different characteristics and requirements [19]. Analytical systems are generally used for the support of specialists and executives. Decisions at these company levels are, in most cases, characterized as strategically or tactically, that means for a longer period. The data access during the execution of analyzes are almost exclusively read-only [19]. On the other hand, transaction systems are used to solve everyday business tasks of a company. In most cases, the time horizon only covers a relatively short period [20]. The typical transactional workload is also largely read access, but compared with analyzes, with a significantly higher

 ISBN: 978-1-61208-571-5 

proportion of write accesses [19]. The merging of OLAP and OLTP systems to an OLXP / HTAP system leads not only to the advantages mentioned, but also to problems and difficulties. From a technical point of view, hybrid workloads (line / column-based & read / write) must be simultaneously processed [21][22][23]. The merging process also leads to a stronger dependency on the respective system provider. In order to be able to exploit the entire benefit of an IMIS, a large number of applications and processes have to be adapted.

IT providers, such as SAP have predominantly driven the hype surrounding the In-memory technology in the past years. The focus of recent developments was mostly technology-oriented. Similar tendencies can be found in early scientific work in this field. Mainly technical features, such as the column-based storage of data [13], data compression [24] or the persistence of volatile storage media [25] were investigated. An alternative approach for the analysis of possible In-memory applications tries to assess the advantages and potentials on the basis of business requirements. In the first papers in this area [26][27] Piller and Hagedorn are investigating factors for evaluating In-memory applications. The authors examine the potential of IMIS in the retail sector. Despite the early maturation phase of this technology at the time of the investigation, initial application patterns have already been identified.

Similar results are also reported by Cundius et. al in their work [28]. They developed a model for evaluating real-time IT systems. The focus of this work was on the workflow-specific properties of real-time IT systems. The use of IMIS not only has an impact on data processing, but also on the downstream decision-making and implementation processes. Vom Brocke et. al examine the connection between the In-memory technology and the resulting business use in their papers [29][30][31]. They conclude that a value-creation for companies is strongly related to the adaption of processes. Vom Brocke et al. as well as Bärenfänger et al. [32] conclude that the introduction of In-memory technology not only leads to a direct benefit, but to a large extent to downstream improvements in the process flow. Meier et al. further pursue the aim of an economic evaluation in [33]. They also divided the economic effects into direct and indirect attributable effects.

One of the most important innovations of IMIS is the combination of analysis and transaction systems. Winter et al. analyze the properties of IMIS in one of the first case studies [34]. In addition to the volume of data, the integration of the analysis and transaction system is identified as the most important indicator for the assessment of IMIS. This point is also highlighted in several other scientific papers in this field [16][26][27][35]. From a solely technical perspective, IMIS offers huge potential. However, the question arises for which companies or application areas this potential can be exploited in practice. For many companies predefined reports and evaluations on a daily basis will still be sufficient. For others, the use of real-time data can become a decisive competitive advantage. Previous application examples often refer to very specific or exotic tasks. A popular example of the application of IMIS is the analysis of sports data, e.g., in Formula 1 [36] or soccer [37]. Although these examples are quite illustrative, they are not suitable to provide insights into the solution of ”everyday” business problems. The lack of economical use cases is regarded as one of the main obstacles to the distribution of IMIS. This is mentioned in science literature [26][35][38] as well as from a company point of view [4][7].

## III. A Framework for the Evaluation and Analyses of In-Memory Based IT Systems

The aim of this research work is to examine and structure IMIS use cases with regard to their success factors. Based on these factors an analysis and evaluation framework has been developed as seen in Figure 1. The methodology described by Klein and Scholl [39] was used to define the overall structure of the framework. The main advantage of the used methodology is the avoidance of structural defects during the modeling phase. Hereby, it is possible to develop a well-designed and feasible decision model. For this purpose, the scope of the model was first restricted in order to consider only the aspects, which are relevant for the problem solving. After the relevant influential factors were identified, they were subdivided through a structural analysis. As a result of this structuring process, an operationalizable target system for assessing and analyzing In-memory use cases has been created. The

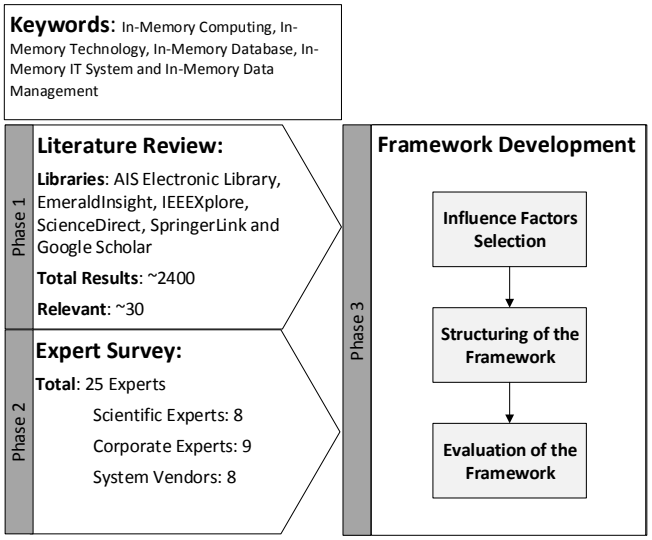



Figure 1. Illustration of the research methodology

methodology of Design Science Research was used as the base for the framework. The advantage of this approach is that the dynamic characteristics of business needs can be directly taken into account. The design process is not static, it allows changes to be incorporated into the existing model [40][41]. In order to gather the basic factors influencing the framework, scientific work and previous case studies in the field of IMIS were analyzed and evaluated during the first design phase. In terms of research method, this was accomplished according to Webster and Watson [42]. In the literature review, established literature databases (AIS Electronic Library, EmeraldInsight, IEEEXplore, ScienceDirect, SpringerLink and Google Scholar) were investigated. The search included the following key words: ”In-Memory Computing”, ”In-Memory Technology”, ”In-Memory Database”, ”In-Memory IT System” and ”In-Memory Data Management”. Subsequently, a backward search was carried out. Therefore only papers dealing with the application and the business perspective of IMIS were used. The study of the literature databases revealed that around 2400 scientific publications have so far dealt with IMIS. Due to the context of this paper only publications with an business perspective were

 ISBN: 978-1-61208-571-5

considered. Hence, 30 relevant papers remained. The detailed results are explained in the next section. During the second design phase,as seen in Figure 1, a qualitative expert survey [43] was carried out to evaluate the results and identify further factors. In particular, the expert survey was carried out to reveal further findings on challenges from an economic point of view. In order to cover a broad range of opinions and experiences, the experts were composed of representatives from different fields. Totally 25 experts in the field of IMIS were interviewed. These included scientists, company representatives as well as representatives of leading IMIS providers. In semi-structured interviews the experts were asked about the potential and the obstacles of the In-memory technology. In addition, the experts were asked to evaluate possible application scenarios and their characteristics in detail. In the final step, the application of the model will be presented using selected In-memory use cases from the retail sector.

### A. Results of the literature and case study review

The examined works used different approaches to deal with the analysis and assessment of the scenarios. The work [26] by Piller and Hagedorn has proven to be a suitable basis for the model presented in this work. Starting from the business process characteristics described in this study, further influencing factors were identified and classified. In the following, the examined factors are presented and explained.

Main memory-based databases are often mentioned to solve the challenges which are associated with so called Big Data applications. Due to the availability of larger main memory and advanced compression by the column orientation, IMIS is able to process large amounts of data [18][44]. Therefore, it is appropriate to include the data volume of a use case into the consideration. Apart from the data volume, a number of other factors play a decisive role. These include, for example, the urgency of the results [26][27][31][45] or the dynamics of the data [26][27][28]. Hence, high-performance systems have a strong positive effect if the data changes frequently. If the underlying data changes only very rarely and to a small extent, the potential additional value of a real-time result is very limited. An example for this are purchase proposals in large online shops based on customer segmentation, which change in general only rarely or marginally. A further influencing parameter is the number and type of source systems [46][47]. In order to cover a broad range of information, it may be advantageous to integrate several different source systems. However, from a critical point of view, problems emerge. The transmission from external sources can lead to delays. A further and currently very often-discussed topic is the veracity of information [48].

As already mentioned in section II, business processes must be adapted with regard to the newly gained flexibility and speed of data analysis in order to exploit the full potential [28][29][30]. The need for process adaptation has to be clarified on the base of the time business-value relationship concept from Hackathorn [45]. Figure 2 visualizes this concept and shows that the information-processing latency caused by IMIS can be reduced, but the additional business profit is relatively low. In order to generate a higher added value, it is also necessary to modify and accelerate the downstream decision-making and implementation processes.

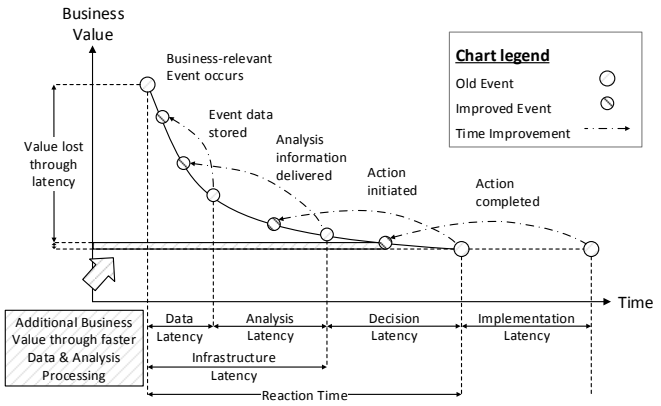



Figure 2. Correlation between time and business value (adapted according to [45])

### B. Results of the expert survey

In order to evaluate the results and identify further influential factors, an expert study was conducted. One of the most frequent mentioned points in the interviews was the uncertain investment security. Despite the decline in hardware costs, the purchase of a main memory-based information system is associated with both high investment costs and a significant total cost of ownership [33]. As with any other investment decision, sufficient value must be generated to cover the cost of acquisition. A large proportion of the interviewed company representatives have criticized the poor cost-benefit ratio concerning IMIS and mentioned several reasons for that evaluation. In most business applications, mainly ”conventional” analyses and evaluations are carried out. These are already defined in advance or can be well predicted and scheduled. Due to the tactical or strategic character of the decisions, there is no exceptional urgency to obtain the results in most cases.

Apart from traditional OLAP tasks, the In-memory technology is perceived more positively. This includes, for example, the areas of predictive maintenance or the integration and analysis of social media. To implement a predictive maintenance, a large number of sensors must be integrated into the analytical system. The continuous measurement results in a high volume of data. Ideally, these data should be analysed as quickly as possible. Another example is the processing of social media, where large quantities of unstructured texts have to be processed. These two examples already confirm a significant proportion of the influencing factors from the first design phase. Another important criterion mentioned frequently were implementation conditions. According to the experts, not only the speed of decision-making is a relevant factor, but also the technical effort and legal obstacles that have to be considered. These factors were not taken into account in the previous literature. Efforts for the indoor localization or digital price tags were cited as examples for technical obstacles. An example for legal obstacles are the data privacy laws regarding the analysis of personal data, especially in EU countries like Germany, Spain or the Netherlands.

### C. Conception and structuring of the framework

The literature review as well as the results from the expert study make clear that a variety of factors influence the assessment of IMIS scenarios. For a systematic analysis, it is necessary to structure the identified influencing factors.

 ISBN: 978-1-61208-571-5

Figure 3 summarizes the results gathered from both the literature review and the expert study. Based on the results of the literature review, the factors can be clustered into two main categories: data and analysis factors. In the category analysis factors, a large part of the investigations dealt with questions of urgency, complexity or flexibility of analysis. Another segment of research focuses on data-driven factors. These include, among other issues, the volume, the topicality and the dynamics of data. As the hesitant spread of IMIS shows, the technical advantages alone are not enough to generate a substantial benefit. In the past research of IMIS, this fact was rarely taken into account. To consider aspects which are related to, e.g., real-time decisions and to take the results of our expert study into account, the categories were extended by the category of economic factors. This category contains factors with regard to internal as well as external implementation conditions, which are particularly important in the corporate context. The different characteristics of the factors show that some have a positive effect on the use of IMIS, while others have a negative impact. This means, e.g., that high requirements regarding the urgency of evaluations have a positive effect on the evaluation of an IMIS. To take this into account, we have added an influence indicator to our framework.

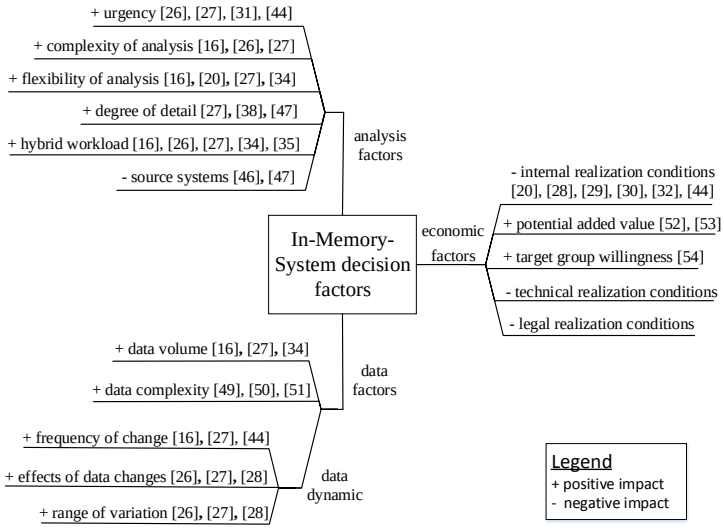



Figure 3. Overview of the analysis and evaluation framework

### D. Application example of the framework

The functionality of the framework will be shown based on selected application examples from the retail sector. These examples were discussed during the expert interviews and in first case studies. For reasons of space, only the example ”analysis of sales and inventory data” [26] is explained in detail. For a better interpretation the results are summarized in the end.

*1) Case study ”analysis of sales and inventory data”.:* The goal of the ”analysis of sales and inventory data” [26] scenario is to discover anomalies in advertising and sales figures. High requirements are formulated regarding the urgency, the volume of data and the dynamics of the data of the IT system. The analysis of the information is mostly based on recurring standard reports. The complexity of the evaluation as well as the complexity of the underlying data in this case study is typically low. Sales documents in the retail sector are well-structured and can therefore be easily processed. A complex transformation is not necessary. All required data can be obtained from internal sources. Hence, there is no latency by loading data from external sources. Due to the high range of fluctuations in sales figures [55], the economic benefit of rapid intervention is quite high. At this point, the question arises which measures can be taken to minimize the fluctuations. In the work of Piller and Hagedorn, only non-price measures to reduce sales fluctuation are proposed [26]. In this case, there are no legal or technical obstacles to the realization of the measures. Evaluating the presented scenario characteristics it becomes clear that analytical requirements are on a high level and the data requirements are on a medium level. Due to the high economic benefits and low implementation obstacles, the overall potential for IMIS according to our framework is considered very high. In contrast to this, other examples show significant differences in the technical as well as the economic indicators.

*2) Case study ”dynamic prices”.:* For the example of dynamic prices [30], similar conditions apply as for the first scenario. From a technical perspective, the determination of dynamic and customer-specific prices is hardly a problem for current IT systems. In order to deploy a price differentiation, further prerequisites must be fulfilled. The information when a customer enters the shop and where he is located has to be available [56]. From a technical point of view, the challenge is to locate the customer exactly on several floors. Another fundamental requirement is the customers willingness to pay that individual price [54]. Reports about the introduction of individual discounts by a German retail company for example, led to extensive criticism and reluctance of the customers [57].

*3) Evaluation of the case study results.:* The evaluation of the attributes from a technical point of view implies the suitability of an IMIS for both use cases. The large data volume, the recently changing data, the need for quick responses and the combination of analytical and transactional tasks are strong indicators. The benefit of the presented framework becomes especially clear when the economic factors are considered. These factors indicate problems regarding the realization of the second example. With the help of the structured model shown in Fig.2, corporate decision makers can create a more holistic evaluation of potential use cases.

## IV. Conclusion

The aim of this work was to create a framework for analyzing and evaluating application scenarios in the context of IMIS. As current research as well as statements from industry experts show, such a framework was missing. To cover all relevant factors for the application of an IMIS, not only theoretical work was included in this work. Through the inclusion of corporate experts, also practical aspects have been considered. Based on some of the first case studies in this area and scientific work, a large part of the influencing factors could be identified. Results show that the influence factors found through literature review and expert study could be divided into three main categories: analysis factors, data-driven factors and economic factors. Based on the expert survey, it was also possible to confirm the factors from the literature and to uncover other previously unconsidered factors. In order to take account of all aspects relevant to the companies, the model was

 ISBN: 978-1-61208-571-5 

expanded by features with regard to the profitability and the feasibility of possible fields of application. These include, for instance, the implementation conditions, legal obstacles or the willingness of target groups. Using the presented model, it is possible to examine potential and existing real-world scenarios regarding their requirements and their feasibility. In a next step, it will be necessary to evaluate the suitability of the framework based on quantitative investigations. To consider branch-specific variations of the relevance of the influence factors it is further useful to include corresponding weighting factors.

## REFERENCES


[1] "Amazon Introduces Prime Now," 2014, URL: http://www.businesswire.com/multimedia/home/20141218005363/en/ [accessed: 2017-03-14].

[2] "The Wall Street Journal - High-Speed Stock Traders Turn to Laser Beams," 2014, URL: http://www.wsj.com/articles/SB10001424052702303947904579340711424615716 [accessed: 2017-03-14].

[3] H. Rosa, Beschleunigung: die Veränderung der Zeitstrukturen in der Moderne, ser. Suhrkamp Taschenbuch Wissenschaft. Suhrkamp, 2005.

[4] "SAP fehlen echte HANA-Business-Cases," 2015, URL: http://www.cio.de/a/sap-fehlen-echte-hana-business-cases,2940526 [accessed: 2017-03-14].

[5] "Lack of SAP HANA use cases stifling demand among ASUG members," 2014, URL: http://diginomica.com/2014/08/08/lack-sap-hana-use-cases-stifling-demand-among-asug-members/ [accessed: 2017-03-14].

[6] "ASUG Member Survey Reveals Successes, Challenges of SAP HANA Adoption," 2014, URL: http://www.asugnews.com/article/asug-member-survey-reveals-successes-challenges-of-sap-hana-adoption [accessed: 2017-03-14].

[7] "Hunting happy HANA customers," 2014, URL: http://diginomica.com/2014/10/27/hunting-happy-hana-customers/ [accessed: 2017-03-14].

[8] "SAP Business Suite powered by SAP HANA," 2014, URL: https://www.pac-online.com/download/9757/125462 [accessed: 2017-03-09].

[9] H. Garcia-Molina and K. Salem, "Main memory database systems: An overview," IEEE Transactions on knowledge and data engineering, vol. 4, no. 6, 1992, pp. 509–516.

[10] D. J. DeWitt, R. H. Katz, F. Olken, L. D. Shapiro, M. R. Stonebraker, and D. A. Wood, Implementation techniques for main memory database systems. ACM, 1984, vol. 14, no. 2.

[11] F. Färber, S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg, and W. Lehner, "Sap hana database: data management for modern business applications," ACM Sigmod Record, vol. 40, no. 4, 2012, pp. 45–51.

[12] M. K. Gupta, V. Verma, and M. S. Verma, "In-memory database systems-a paradigm shift," 2014, pp. 333–336.

[13] D. J. Abadi, P. A. Boncz, and S. Harizopoulos, "Column-oriented database systems," Proceedings of the VLDB Endowment, vol. 2, no. 2, 2009, pp. 1664–1665.

[14] A. Kemper and T. Neumann, "Hyper: A hybrid oltp&olap main memory database system based on virtual memory snapshots," in Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, 2011, pp. 195–206.

[15] H. Plattner, "A common database approach for oltp and olap using an in-memory column database," in Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009, pp. 1–2.

[16] P. Loos, J. Lechtenbörger, G. Vossen, A. Zeier, J. Krüger, J. Müller, W. Lehner, D. Kossmann, B. Fabian, O. Günther et al., "In-memory-datenmanagement in betrieblichen anwendungssystemen," Wirtschaftsinformatik, vol. 53, no. 6, 2011, pp. 383–390.

[17] M. Pezzini, D. Feinberg, N. Rayner, and R. Edjlali, "Hybrid transaction/analytical processing will foster opportunities for dramatic business innovation," Gartner (2014, January 28) Available at https://www. gartner. com/doc/2657815/hybrid-transactionanalytical-processing-foster-opportunities, 2014.

[18] H. Plattner and A. Zeier, In-memory data management: technology and applications. Springer Science & Business Media, 2012.

[19] J. Krueger, M. Grund, C. Tinnefeld, B. Eckart, A. Zeier, and H. Plattner, "Hauptspeicherdatenbanken für unternehmensanwendungen," Datenbank-Spektrum, vol. 10, no. 3, 2010, pp. 143–158.

[20] S. Strohmeier, "Hauptspeicherdatenbanken in der betrieblichen informationsversorgung–technische innovation und fachliche stagnation," Wirtschaftsinformatik, vol. 54, no. 4, 2012, pp. 209–210.

[21] J. Krueger, M. Grund, C. Tinnefeld, H. Plattner, A. Zeier, and F. Faerber, "Optimizing write performance for read optimized databases," in Database Systems for Advanced Applications. Springer, 2010, pp. 291–305.

[22] J. Wust, J.-H. Boese, F. Renkes, S. Blessing, J. Krueger, and H. Plattner, "Efficient logging for enterprise workloads on column-oriented in-memory databases," in Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012, pp. 2085–2089.

[23] A. Kemper, T. Neumann, F. Funke, V. Leis, and H. Mühe, "Hyper: Adapting columnar main-memory data management for transactional and query processing." IEEE Data Eng. Bull., vol. 35, no. 1, 2012, pp. 46–51.

[24] D. Abadi, S. Madden, and M. Ferreira, "Integrating compression and execution in column-oriented database systems," in Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006, pp. 671–682.

[25] T. Winsemann and V. Köppen, "Kriterien für datenpersistenz bei enterprise data warehouse systemen auf in-memory datenbanken." in Grundlagen von Datenbanken, 2011, pp. 97–102.

[26] G. Piller and J. Hagedorn, "In-memory data management im einzelhandel: Einsatzbereiche und nutzenpotentiale," Multikonferenz Wirtschaftsinformatik 2012 : Tagungsband der MKWI 2012 / Hrsg.: Dirk Christian Mattfeld; Susanne Robra-Bissantz.

[27] ——, "Business benefits and application capabilities enabled by in-memory data management," in Innovative Unternehmensanwendungen mit In-Memory Data Management, IMDM 2011, 2. Dec 2011, Mainz, ser. LNI, W. Lehner and G. Piller, Eds., vol. 193. GI, 2011, pp. 45–56.

[28] C. Cundius and R. Alt, "Real-time or near real-time?-towards a real-time assessment model," Thirty Fourth International Conference on Information Systems, 2013, pp. 1–18.

[29] J. vom Brocke, S. Debortoli, and O. Müller, "In-memory database business value," 360 - The Business Transformation Journal, vol. 3, no. 7, 2013, pp. 16–26.

[30] J. vom Brocke, "In-memory value creation, or now that we found love, what are we gonna do with it?" BPTrends, vol. 10, 2013, pp. 1–8.

[31] J. vom Brocke, S. Debortoli, O. Müller, and N. Reuter, "How in-memory technology can create business value: insights from the hilti case," Communications of the Association for Information Systems, vol. 34, no. 1, 2014, pp. 151–167.

[32] R. Bärenfänger, B. Otto, and H. Österle, "Business value of in-memory technology–multiple-case study insights," Industrial Management & Data Systems, vol. 114, no. 9, 2014, pp. 1396–1414.

[33] M. C. Meier, A. Scheffler, and K. Finanz, "Ökonomisch sinnhafte bewertung von in-memory-basierten betrieblichen informationssystemen." in IMDM. Citeseer, 2011, pp. 115–124.

[34] R. Winter, S. Bischoff, and F. Wortmann, "Revolution or evolution? reflections on in-memory appliances from an enterprise information logistics perspective." in IMDM, 2011, pp. 23–34.

[35] R. Schütte, "Analyse des einsatzpotenzials von in-memory-technologien in handelsinformationssystemen." in IMDM, 2011, pp. 1–12.

[36] "McLaren Formula 1 - Partners - SAP," 2015, URL: http://www.mclaren.com/formula1/partners/SAP [accessed: 2017-03-14].

[37] "Big Data & Spatial Analytics Help Germany Score the World Cup," 2014, URL: http://www.saphana.com/community/blogs/blog/2014/07/15/how-big-data-helped-germany-score-the-world-cup [accessed: 2017-03-14].


 ISBN: 978-1-61208-571-5 

[38] G. Koleva, “Fields of usage for in-memory databases in enterprises,” in 11th Workshop on Information Systems and Services Sciences, 2011, pp. 19–33.

[39] R. Klein and A. Scholl, Planung und Entscheidung. Vahlen, München, 2004.

[40] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” MIS quarterly, vol. 28, no. 1, 2004, pp. 75–105.

[41] A. R. Hevner, “A three cycle view of design science research,” Scandinavian journal of information systems, vol. 19, no. 2, 2007, p. 4.

[42] J. Webster and R. T. Watson, “Analyzing the past to prepare for the future: Writing a literature review,” MIS quarterly, 2002, pp. xiii–xxiii.

[43] H. O. Mayer, Interview und schriftliche Befragung: Grundlagen und Methoden empirischer Sozialforschung. Walter de Gruyter, 2013.

[44] H. Zhang, G. Chen, B. C. Ooi, K.-L. Tan, and M. Zhang, “In-memory big data management and processing: A survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 7, 2015, pp. 1920–1948.

[45] R. Hackathorn, “Minimizing action distance,” DM REVIEW, vol. 12, 2002, pp. 22–23.

[46] M. Nadj and C. Schieder, “Quo vadis real-time business intelligence? a descriptive literature review and future directions,” 24th European Conference on Information Systems, 2016, pp. 1–20.

[47] W. H. Inmon, Building the data warehouse. John wiley & sons, 2005.

[48] L. Berti-Equille and J. Borge-Holthoefer, “Veracity of data: From truth discovery computation algorithms to models of misinformation dynamics,” Synthesis Lectures on Data Management, vol. 7, no. 3, 2015, pp. 1–155.

[49] H. Baars and H.-G. Kemper, “Management support with structured and unstructured dataan integrated business intelligence framework,” Information Systems Management, vol. 25, no. 2, 2008, pp. 132–148.

[50] B. Inmon and K. Krishnan, Building the Unstructured Data Warehouse: Architecture, Analysis, and Design. Technics Publications, 2011.

[51] S. Sarawagi, “Queries over unstructured data: Probabilistic methods to the rescue,” in International Workshop on Business Intelligence for the Real-Time Enterprise. Springer, 2009, pp. 1–13.

[52] L. Olsson and C. Janiesch, “Real-time business intelligence und action distance: Ein konzeptionelles framework zur auswahl von bi-software.” in Wirtschaftsinformatik, 2015, pp. 691–705.

[53] R. Hackathorn, “Real-time to real-value,” Information Management, vol. 14, no. 1, 2004, p. 24.

[54] M. Bauer, Kundenzufriedenheit in industriellen Geschäftsbeziehungen: kritische Ereignisse, nichtlineare Zufriedenheitsbildung und Zufriedenheitsdynamik. Springer-Verlag, 2013.

[55] C. Narasimhan, S. A. Neslin, and S. K. Sen, “Promotional elasticities and category characteristics,” The Journal of Marketing, 1996, pp. 17–30.

[56] “Why indoor navigation is so hard,” 2011, URL: http://radar.oreilly.com/2011/10/indoor-navigation.html [accessed: 2017-03-14].

[57] “Jeder hat seinen Preis,” 2014, URL: http://www.zeit.de/wirtschaft/2014-10/absolute-preisdiskriminierung [accessed: 2017-03-15].

 ISBN: 978-1-61208-571-5 

# Smartphone-based Data Collection with Stunner Using Crowdsourcing: Lessons Learnt while Cleaning the Data

Zoltán Szabó*, Vilmos Bilicki*, Árpád Berta†, and Zoltán Richárd Jánki*
*Department of Software Engineering
University of Szeged, Hungary
Email: {szaboz, bilickiv, jankiz}@inf.u-szeged.hu
†MTA-SZTE Research Group on AI
University of Szeged, Hungary
Email: berta@inf.u-szeged.hu

***Abstract*—The increasing popularity of smartphones makes them popular tools for various big data collecting crowdsourcing campaigns, but there are still many open questions about the proper methodology of these campaigns. Beyond this, despite the growing popularity of this type of research, there are familiar difficulties and challenges in handling a wide range of uploads, maintaining the quality of the datasets, cleaning the data sets containing noisy, incorrect data, motivating the participants, and providing support for data collecting regardless of the remoteness of the device. In order to collect information about the Network Address Translation (NAT) related environment of mobile phones, we utilized a crowdsourcing approach. We collected more than 70 million data records from over 100 countries measuring the NAT characteristics of more than 1300 carriers and over 35000 WiFi environments during the three year project. Here, we introduce our data collecting architecture, some of the most prominent problems we have encountered since its launch, some of the solutions and proposed solutions to handle difficulties.**

***Keywords–smartphones; data cleaning; crowdsourcing.***

## I. MOTIVATION

In recent years, smartphones have become part of our everyday lives. Their wide range of uses along with multiple sensors, networking and computational capabilities have also made them seemingly ideal platforms for research. One research area is data collection, with the collected datasets available for a wide area of analysis, including network mapping, discovering and analyzing various networks, and the network coverage of certain areas.

Different research teams from all over the world have discovered these new opportunities, and they employ smartphones as crowdsourcing tools in a wide variety of ways. Through crowdsourcing, they assign tasks to different users with different device types to collect data in real-life situations, or a monitored environment, providing huge amounts of realistic data. In recent years, we have seen a lot of successful, and interesting approaches to this methodology.

However, it is still a question of how exactly crowdsourcing campaigns should be implemented. Several research projects, such as SmartLab [1], the behaviour-based malware detection system Crowdroid [2], and the cross-space public information crowdsensing system FlierMeet [3] recruited a small number of users, who could be trusted, contacted if necessary, and provided the data taken from a known environment, specifically chosen, or created for the crowdsourcing project. This limited the variability and the amount of the data, but the results were of a high quality and easy to validate.

Another approach for recruitment is to upload the smartphone application to the Google Play store, or the Apple App Store, making it available for download by anyone world-wide, and opening up data collecting opportunities for anyone who agreed to the terms and services of the software package. With proper marketing, the results could include enormous datasets obtained from around the world. The NoiseTube project [4] for crowdsourcing noise pollution detection was downloaded by over 500 people from over 400 regions world-wide. The Dialäkt App [5], one of the most well-known crowdsourcing campaigns in recent years, was the most downloaded iPhone app in Switzerland after its launch, with wide media coverage, and over 78000 downloads from 58923 users by the time they had published their results. Many more datasets were collected in the Bredbandskollen project, later to be used by various smartphone-based research projects [6], which has collected network data from 3000 different devices and over 120 million records since its launch in 2007, and the OpenSignal [7] application, which between 2012 and 2013 collected over 220 million data records from more than 530000 devices and from over 200 countries.

However, collecting data using smartphones is not without its difficulties, and there are a number of challenges when smartphones are used as the prime source of information. Among these, battery consumption and network state are among the most important elements, as constantly accessing the state of the phone sensors and listening to specific events takes a heavy toll on the battery, making data collection inadvisable in certain situations (for example, after a device signalled a battery low event), and it is feasible, but pointless in other situations (the phone is on a charger while the user is asleep - the energy is there, but the valuable information is only a fraction of what we would get from an active user). Network state again has to be taken into account, as even today in many environments, we cannot ensure that a device will always have a connection strong enough to send the collected data to the server. Privacy is also an issue, since the data has to remain identifiable yet not contain any trace of personal information.

Aside from all of the above, if data collection was successful, we still have the problem of noisy, incorrect, disorganized data. We have to take into consideration the fact that there are different devices, different versions of the same OS, bugs, such as duplicated records uploaded by the client on network error and damaged records resulting from a similar event. We also have to take user interference into account, who may not wish to provide valuable data (e.g., by deliberately leaving the phone at home on a charger, having it switched off during specific hours, etc.). Their results will still be counted as valuable data, but this can severely distort the collected data set, as well as the results used in evaluations.

An even bigger problem, when crowdsourcing is a global campaign, is that of time synchronization. Not only do we have to find a good solution for the various time zones of the devices, but also the different time codes of the phone collecting and sending the data and the server storing this data, and the possibility that the user might have manually altered the date and time on the test phone as well.

All of the possibilities mentioned above result in a mixed

 ISBN: 978-1-61208-571-5 

situation where the power and potential of the smartphones, as research tools cannot be denied, but to acquire correct, useful data is a challenge in itself. This requires careful planning, taking into account almost every possible cause of data distortion, well-defined filters and data cleaning algorithms before any actual research can be performed on the data collected. In this article, we are going to present our solutions with this type of data collection, and our solutions to the problems that emerged.

Our goal was to develop an Android app in order to collect important network information for research on the peer-to-peer (p2p) capabilities of smartphones, including the NAT type, network type and network provider. It does so by taking measurements on a regular basis as well as during specific events. When taking the measurements, the app sends a request to a randomly chosen Session Traversal Utilities for NAT (STUN) Server from a list, displaying useful network information, such as the IP address and NAT type to the user, while also storing the necessary data in an SQLite database, which later gets uploaded to a data collector server for analysis. The application called Stunner has been available for download from the Play Store since December 2013 [8].

Using the collected data, we will be able to define the graph model of a worldwide, peer-to-peer smartphone network. In this model, we aim to test various peer-to-peer protocols to measure the capabilities of a serverless network architecture, where the phones can slowly update their datasets and generate various statistics, without the data ever leaving this smartphone network. The ultimate goal of our research is the creation of an Application Programming Interface (API), through which developers can utilize these peer-to-peer capabilities to create various data collecting and processing applications (for example, general mood or health statistic researching applications for a specific region) without the need of a processing server.

## II. Literature Overview

The challenges outlined above have been collected from the research results of other teams (Table I) - and nearly all of them offered good design viewpoints during the development of our own data collecting application.

Perhaps the best overview of the possible difficulties was provided by Earl Oliver [9]. While developing a data collecting application for BlueBerry, he defined five of the most common and serious problems, namely volatile file systems on mobile devices (as file systems can be easily mounted and unmounted on nearly any device), the energy constraints, the intervention of third-party applications running in the background, the non-linear time characteristics of the devices, and malicious user activity (file manipulation, simulated manipulation, etc.)

He solved these by exploiting many trends of BlackBerry users: the general maintenance of high battery levels, retrieving manifests of active applications, and data analysis for patterns of manipulation attacks. However, even he could not define a general solution for every problem, and these problems were not the only ones encountered by other research teams. In fact, they found other challenges to be rather common among data collecting applications.

The researchers at Rice University, while developing LiveLab [10], a methodology used to measure smartphone users with a similar logging technique, encountered the problem of energy constraints, with various optimizations needed to lower the high consumption of the logging application. They also recognized the problem associated with data uploading, namely the connectivity to the server which collects the data from the devices and updates them with new information. They chose rsync for its ability to robustly upload any measurement archive which failed earlier.

A similar method of re-uploading the failed archives was used by a research team at the University of Cambridge in their Device Analyzer project [11], which sought to build a dataset that captured real-world usage of Android smartphones, again with a similar event logging based solution. They found that repeated attempts at uploading caused duplicated data on the server, which could simply be removed by the server before saving it to a database. They also solved the above-mentioned problem of nonlinear time by timestamping every measurement with the device's uptime in milliseconds, recording the wall-clock time of the device when their application started, and later recording every adjustment to it by listening to the notifications caused by these adjustments. From these three elements, a simple server-side processing algorithm was able to reconstruct the exact wall-clock time of any given measurement.

Members of the Italian National Research Council [12] also confirmed these challenges (i.e., the scarcity of resources, difficulties with network monitoring and privacy) while also highlighting two more problems, caused by the participants using the devices - the much more complex control tasks in these types of research projects, and the issue of user motivation to get them carry out the tasks required to get valid data.

TABLE I. Comparison of Data Collecting Projects

| *Problem* | *BlackBerry logger* | *Device Analyzer* | *Portolan* | *Livelab* |
|---|---|---|---|---|
| Energy constraints | OS callback based logging | Only 2% of the energy consumption | Computational and analyzation processes are run by the server and the collecting is not too energy consuming | The logging events are optimized, some of the data being collected directly from the system logs |
| Non-linear time | Dates are logged in a UTC timezone; datetime modifications recorded | Every measurement stamped with a device uptime in milliseconds; on startup, the device time is logged, like every modification on device time | Not described (there is a strict communication between client and server, probably kept in sync by this procedure) | Not described, the datetime is most likely to be among the logged data |
| Offline state, unsuccessful upload | - | Batched uploads only when the device is online and the charger is connected | Uploads are handled by proxy servers | Rsync protocol keeps trying until the upload is successful |
| Multiplicated data | - | Every device has a file on the server, multiple copies of data being detected by the server | This is solved by proxy servers | Not described most likely to be filtered by the server |

In this article, we present our experiences with crowdsourcing-based data collection along with our methods and results of data cleaning on the present dataset.

- We propose a solution for the biggest challenge of the batched data uploads, namely the time synchronization among the different elements of the architecture, utilizing a 3-way logging solution, and lightweight log synchronization.
- We introduce heuristics to analyze incorrect NAT values, in order to decide which cases failed because of server side problems, and which cases originated from the client side.
- We also introduce a data cleaning algorithm to correct timestamp overlaps, using battery-based smartphone heuristics to detect anomalies among consecutive measurements, such as excessively rapid charging, or charging when the smartphone is in a discharging state or when no charger is connected.

 ISBN: 978-1-61208-571-5 

## III. OUR FRAMEWORK

### A. Architecture

Our goal with crowdsourcing measurements was to collect information about the network environment of mobile phones. This information is of key importance if one wishes to build a p2p network mobile phones. The crowdsourcing architecture consists of an Android based mobile app, several publicly available STUN servers and a data collecting server. The application collects the information described in Table II.

TABLE II. DISCOVERYDTO OBJECT

| DiscoveryDTO | |
|---|---|
| batteryDTO | Energy supply data specified in the BatteryInfoDTO. |
| wifiDTO | WiFi connection data at the moment of measurement specified in the WifiIntoDTO. |
| mobileDTO | Mobile network data at the moment of measurement specified in the MobileNetInfoDTO. |
| publicIP | The public (external) IP address. |
| localIP | The local (internal) IP address. |
| timestamp | UNIX timestamp at the end of the measurement. |
| androidVersion | The version of Android running on the device. |
| discoveryResultCode | The result of the NAT measurement, defined by the DiscoveryResult enumeration. |
| connectionMode | The connection code used while taking the measurement is defined by the ConnectionType enumeration. |
| triggerCode | The event that triggered the measurement is defined by the DiscoveryTriggerEvent enumeration. |
| appVersion | The version of the application. |
| timeZoneUTCOffset | The difference between UTC and the device time in signed integer format. |

Taking a measurement can be triggered by the events defined in the DiscoveryTriggerEvents enumeration. The event that triggered the measurement lives until the last running test is complete. The enumeration consists of the following items.

- USER: The user started the measurement using the user interface (UI).
- CONNECTION_CHANGED: The broadcast sent by the Android indicated changes in the connection and it triggered the taking of the measurement.
- BATTERY_LOW: The broadcast sent by the Android indicated a low charge level and it triggered the taking of the measurement.
- BATTERY_POWER_CONNECTED: The broadcast sent by the Android indicated a connection to a power supply and it triggered the taking of the measurement.
- BATTERY_POWER_DISCONNECTED: The broadcast sent by the Android indicated a disconnection from a power supply and it triggered the taking of the measurement.
- BATTERY_SCHEDULED: The scheduled battery status control triggered the taking of the measurement, which occurs every 10 minutes.
- BOOT_OR_FIRST_START: Taking the measurement was triggered by the first execution of the application or by the booting of the device.

A measurement starts with an Intent object. The Intent establishes a new DiscoveryService that uses the DiscoveryThreadHandler to start a new thread. The application uses a service implementation running in the background, which may be executed in parallel (with more threads) (Figure 1).

There are two different types of collection, namely online and offline. It is online if there is an active Internet connection on a WiFi or Mobile Network - in this case all the data types mentioned above can be measured and collected. In the case of no Internet connection, there is no guarantee that the network information will be initialized. The schedule in the offline case is set to 30 minutes.

The uploading process occurs when a device is online and it contains at least 10 non-uploaded measurements stored in the local database. After a successful uploading, the records get deleted from the local database in order to avoid duplication. The application did not store any index associated with the measurements, the only field available for this being the timestamp, stored with the discovery data. The storage time is very limited. In fact, after the very first measurement on a given day, the application deletes every record from previous days.

### B. Statistics

Our application went live on 20th December, 2013. To promote the usage of the application we also launched a campaign, during which we provided 80 university students and users with smartphones, who agreed to download and provide data with the application for the duration of one year. On 21st March, 2017 the application had been downloaded and installed by 14,727 users on 745 different device types representing 1300+ carriers and 35000+ WiFi networks.

Although we have not released a new version since 5th January, 2016, the average daily installs have remained unchanged in recent months (the most popular phase being in 2015).

Our target API level was originally 19 (Android 4.4), but the application is still being downloaded and installed on more and newer devices, with Android 6.0 being currently the most popular on active installs.

We have also reached a wide variety of different types of devices, upon which the application got installed. The application also successfully reached hundreds of different mobile providers in different countries, which provided us with various, realistic NAT patterns and traces - which will be important later on, after the data cleaning phase is finished, and the analysis and usage of the data collected has commenced.

*1) The collected data and the most important descriptive statistics:* Based on the size of the dataset collected and our good track record since the 2013 release, it is safe to say that our application and data collecting campaign were both a success (with 70+ million records).

The chart (Figure 2) shows the number of uploaded records per device. The majority of users did not provide any measurements, but the decline of the slope lessened, indicating that the users who provided data were more likely to stay and keep providing data.

During the summer of 2015 we had to reassign some resources to other projects, resulting in an absence of data in the given time period. However, after restarting the server, our input declined only slightly, resulting in a steady amount of data arriving to this day despite the gap of a few months (Figure 3).

An interesting aspect is that although the daily uploads have been pretty steady since the hiatus, the number of active devices providing the uploads has been on a steady decline since early 2016. We hope that with our current developments, this decline can be reversed, and a new record in both the number of active devices and daily uploads obtained (Figure 4).

 ISBN: 978-1-61208-571-5 

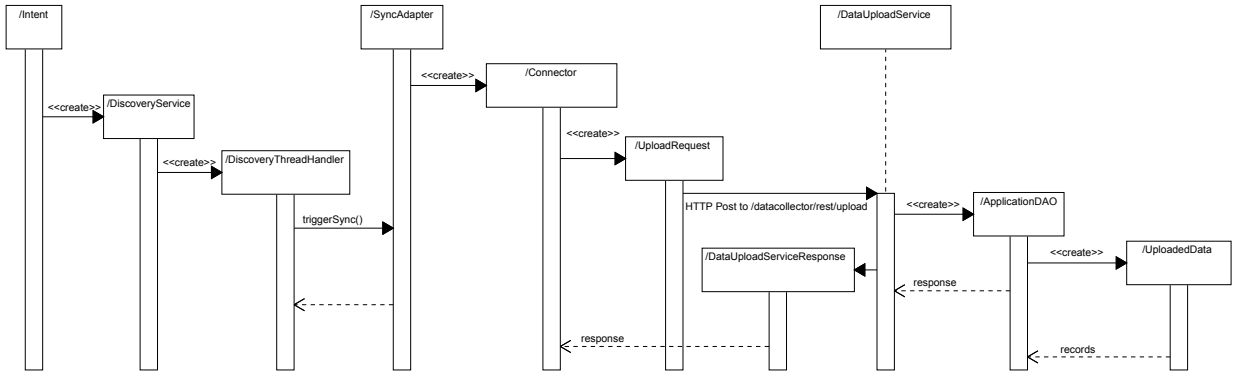



Figure 1. Upload process

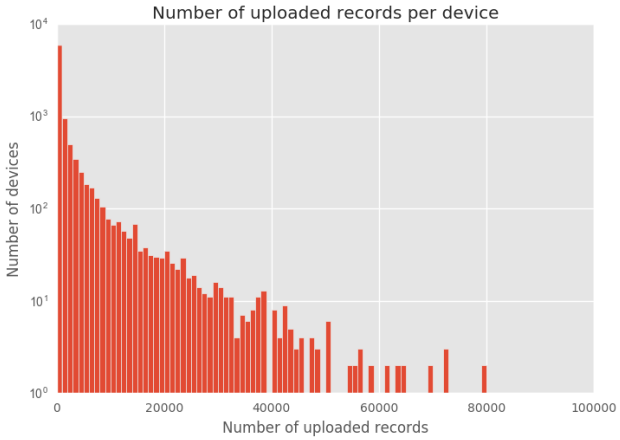



Figure 2. The plot shows the uploaded data per device

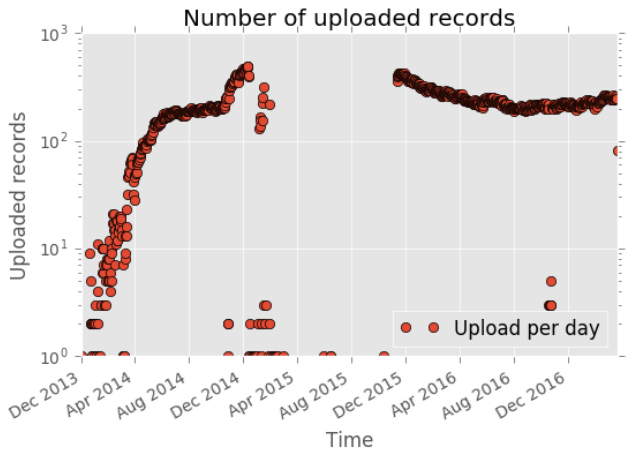



Figure 3. The plot show the uploaded data per day during the whole measurement period.

Following the hiatus, the first spike above shows all the collected data uploaded to the server at the same time. While the Wifi based tests closely follow the trends of active devices and daily uploads, the Mobile Operator-based measurements have been taken at a relatively low, but steady rate (Figure 5). The plot indicates that we can monitor more than 200 mobile networks and over 500 WiFi environments day after day. Here we have shown that a significant amount of data has been collected over the three year period. The real value of the data depends on the quality of the timestamps. Now, we will describe our findings in the area of data cleaning.

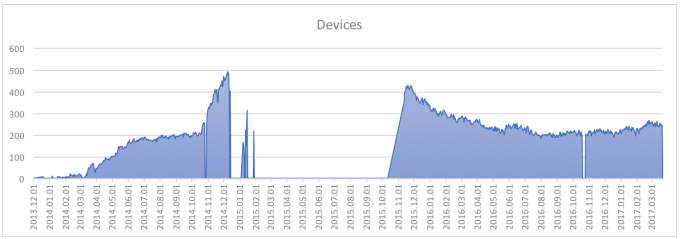



Figure 4. The plot shows the number of active devices per day.

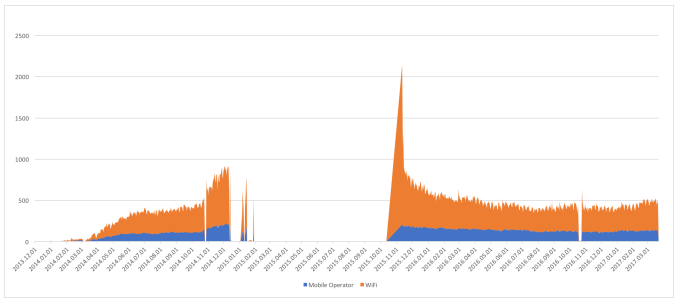


Figure 5. The measurements based on Mobile Operator and WiFi.

## IV. Issues with Collected Data

### A. Data Duplication

In spite of the theoretically sound software environment where the server-side logic was implemented in JEE with transactional integrity taken into account, it turned out that a significant proportional percentage of the dataset had been duplicated. We applied simple heuristics in order to filter out the duplicate measurement records by comparing only the client-side content and skipping the server-side timestamp and other added information. In practice, we utilized the Python Pandas framework duplicate filtering method shown in Figure 6, to remove the duplicates.

We found that out of the 70+ million rows only 30+ million rows were unique, while the remaining part were duplicates. We investigated the possible root cause of this phenomenon. Figure 7 shows the total submitted records per device versus the duplicated records per device. It clearly shows that there is a linear relationship between the two values. This is an evidence that this is a system-level symptom and not a temporal one related to the server overloading. The same is true if we check the temporal dimension of the duplicated

PSEUDO CODE:
```
data['duplicated'] = = data.duplicated(subset=[all client side columns] , keep='first')
```

Figure 6. JSON sample

 ISBN: 978-1-61208-571-5

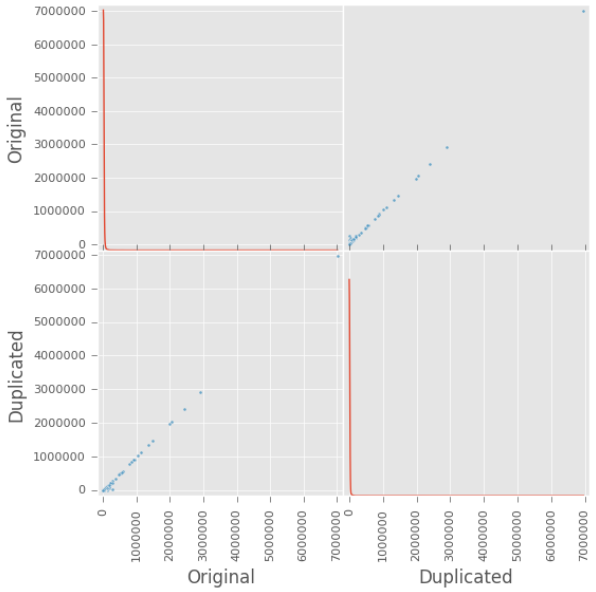



Figure 7. The plot shows the duplicated data per device vs the total number of records submitted by that device

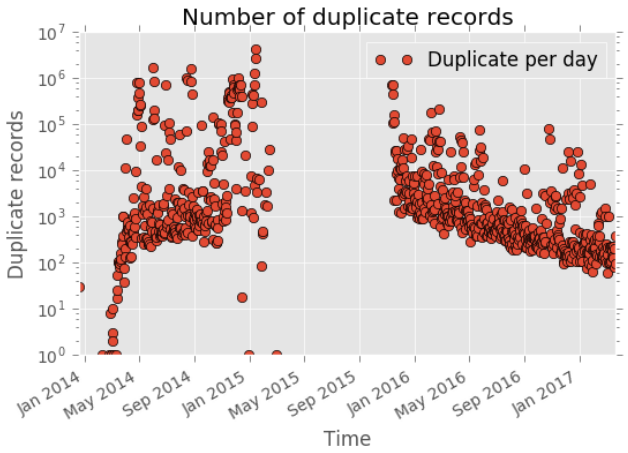



Figure 8. The plot concerning the duplicated data per day (server side)

records during the given period (Figure 8).

After an in-depth investigation of the client code, we found that the default HttpClient configuration contained a very robust upload model, with a default value of 3 retries for every HyperText Transfer Protocol (http) operation, if it failed with a timeout. This is a very useful method for simple data upload, but in our case, if the timeout chosen in the settings was too short, the client might have uploaded the same batch of records up to four times to the server, which would acknowledge and store all of them. In order to stop further multiple uploads, we will need to carefully look at the correct timeout and retry values and also identify the upload batches, so the server will be able to detect the retries on upload.

*1) Overlap of the client-side timestamps:* The actual unreliability of the client-side timestamps was a surprise for us. Figure 9 shows the difference between the Android timestamp and the date captured on the server side. A significant number of measurements have big differences between these two dates. The difference between the two timestamps is only an indication that there could be an error in the measurements as a week or weeks may pass by after capturing and uploading

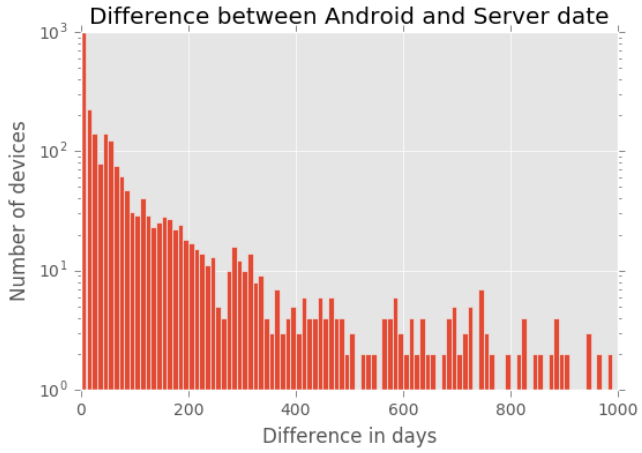



Figure 9. Difference between the Android and Server date

TABLE III. HEURISTICS FOR DETECTION

| *Name* | *Description* | *Detection capability* |
|---|---|---|
| Fast change detector (ABC and SBC) (the first letter codes the ordering applied: A - android, S - server side, this coding being consistent among the different detectors) | We used the battery percentage and its sluggish behaviour to detect the fast changes. We defined the speed of change as the ratio of the two consecutive timestamps and the battery percentage difference between these two timestamps. We defined a threshold high enough to be able to recognize the measurement as an error. | For time-reset starting date estimation. |
| Rules based on charging and plugged state | This method focuses on the rules defined without time being included.<br>Rules: Charging (more than 20% change) while not on charger (ACEU-SCEU)<br>Charging (more than 20% change) while in discharging state (ACED-SCED)<br>Big changes between consecutive elements (charging 20%, discharging 6%) (AP-SP)<br>Charging (more than 20% change) while not on the charger and the phone is in a discharging state (ASC-SSC) | These methods could be applied in order to detect the beginning of a new measurement period (among the overlapped timestamps) |

the data to the server in the case of missing or inadequate network conditions.

We started to examine the nature of the Android timestamp. First, we noticed records with timestamps that were significantly earlier (e.g., 01.01.1970) or later (01.01.2023) than our other measurements. Finding invalid time periods was trivial (like 2023), but it transpired during our in-depth investigation that several phones were reset to a valid date that lay within the observation period. In order to be able to properly detect this anomaly, we elaborated several simple heuristics for detection, these being shown in Table III.

We applied the anomaly detection heuristics mentioned above in order to compare two basic sorting approaches; namely, sorting by the server-side information (e.g., serial number) and the sorting based on the mobile timestamp. We observe that for about 6-7 thousand devices the number of errors is zero. So about 1/8 of the total devices are affected by the time overlap. Figure 10 shows the results of the fast change detector applied for the two ordering approach (it was run on a filtered dataset, skipping the valid data). The green line (server-side sorting) indicates fewer fast change errors in most cases (it was able to eliminate this error on about 40% of the affected devices). The scatterplot below (Figure 11) also shows a clear correlation between the two sorting approaches and the number of fast change errors. The slope of the correlation line (and the points under the line) tells us that the server side sorting was able to reduce the fast change errors in most cases. Based on these findings, one simple approach for time overlap fixing might be the hybrid sorting approach where a given number of records are located after a fast change error had been sorted

 ISBN: 978-1-61208-571-5 


Figure 10. Fast change errors

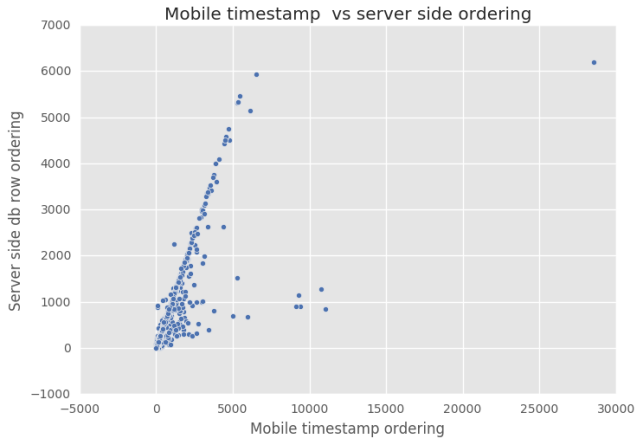



Figure 11. Mobile timestamp and server-side ordering

based on the server-side sorting.

The effectiveness of the simple server-side sorting is also shown in the Figure 12 concerning the correlation between different error detection heuristics and the sorting methods. It is apparent that server-side sorting can significantly decrease the error level for all error detectors (when comparing the same method with S and A sorting, most of the points are below or above the similarity line).

With the previously described heuristics, we were able to demonstrate that the server-side sort order can reduce the rows suspected of being in the wrong position to about 1/10 of the total dataset. A further decrease in the suspected errors could be achieved with a richer ruleset that incorporated different mathematical models for batteries. For our purposes, the current reliability level of the causality dimension of the data set is quite sufficient.

### *B. NAT discovery result code corrections*

The main feature of our application is the discovery of the NAT type. Users can ask the application about their NAT information and public IP address. This method is based on User Datagram Protocol (UDP) message-based communication between the device and a randomly picked STUN server. A STUN server can discover the public IP address and the type of NAT that the clients are behind.

We were faced with a problem that was caused by the prefixed STUN server list. It contains a list of 12 reliable

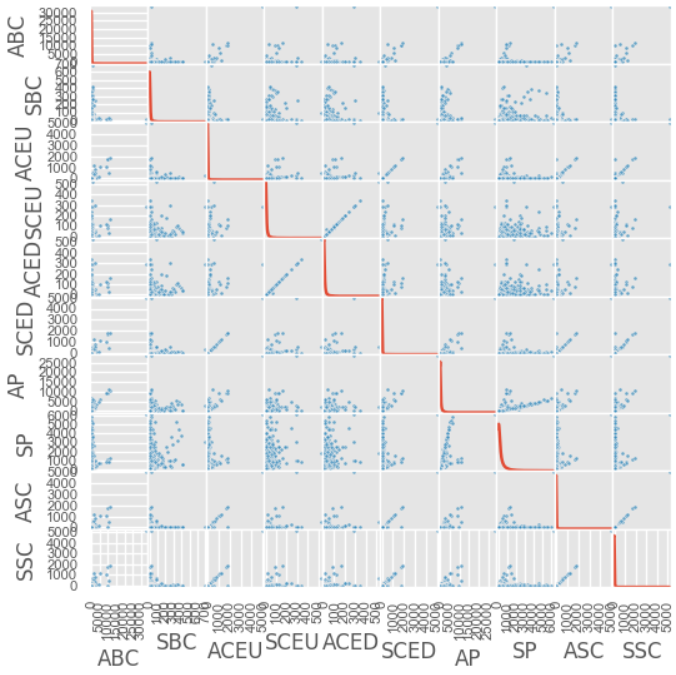



Figure 12. Error detectors

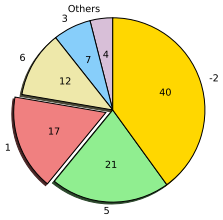
servers that are suitable for NAT detection, this list being embedded inside the application code. It allows the device to randomly pick a STUN server. As a result, every measured NAT type in the timeline is based on a different STUN server's NAT test. Hence it makes the measured data more trustworthy. This random pick approach has been well designed and worked very well initially. However, after a time four of the STUN servers went offline without any prior notice. Since then this four failed STUN server provide the same NAT discovery result code as firewall blocked connections. As a result of this error, some uncertainty exists in the NAT discovery result code. Therefore we propose a solution on how to correct it and make the collected data useable afterward. Quite significantly, another solution is needed to avoid connections to a failed STUN server.

Now we need to discuss the obscure NAT discovery result code. This is the 16.76% of the total measurement records. We have carried out this examination over the dataset, which has already been prefiltered and processed, the order being based on the approach defined above (Figure 13).

Firstly we need to discuss the FIREWALL_BLOCKS result code. This code is corresponding to NAT tests that has open communication channel but never get response from the STUN server. In normal case it means that firewall blocks the connection. Unfortunately records also has no response from failed server. Even though a part of those records may have online NAT type. Therefore these records are uncertain and further examination is needed.

Below, we present a method for filter the STUN server errors from FIREWALL_BLOCKS discovery result code. These set of records contains uncertain potential online states. The server fails with a 4/12 probability, and the event of consecutive repeated fails has an exponential pattern. Consequently we define sessions with consecutive repeated FIREWALL_BLOCKS discovery result codes and look at their distribution. If the distribution is roughly an exponential distribution, then we can interpret them as online and we can define their network properties. Otherwise, the others that do

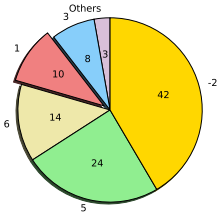
 ISBN: 978-1-61208-571-5 




*NAT Discovery Result Code*

| | | | |
|---|---|---|---|
| -2 | NO_CONNECTION | 11951611 | 40.03% |
| -1 | ERROR | 29 | < 0.01% |
| 0 | OPEN_ACCESS | 706662 | 2.37% |
| 1 | FIREWALL_BLOCKS | 5002770 | 16.76% |
| 2 | SYMMETRIC_FIREWALL | 198442 | 0.66% |
| 3 | FULL_CONE | 1993739 | 6.68% |
| 4 | RESTRICTED_CONE | 280969 | 0.94% |
| 5 | PORT_RESTRICTED_CONE | 6240586 | 20.90% |
| 6 | SYMMETRIC_CONE | 3482591 | 11.66% |

Figure 13. Discovery Result Code




*NAT Discovery Result Code after correction*

| | | | |
|---|---|---|---|
| -2 | NO_CONNECTION | 12401805 | 41.53% |
| -1 | ERROR | 29 | < 0.01% |
| 0 | OPEN_ACCESS | 284206 | 0.95% |
| 1 | FIREWALL_BLOCKS | 2999985 | 10.05% |
| 2 | SYMMETRIC_FIREWALL | 230693 | 0.77% |
| 3 | FULL_CONE | 2295290 | 7.69% |
| 4 | RESTRICTED_CONE | 318136 | 1.07% |
| 5 | PORT_RESTRICTED_CONE | 7272018 | 24.35% |
| 6 | SYMMETRIC_CONE | 4055237 | 13.58% |

Figure 14. NAT Discovery Result Code after correction

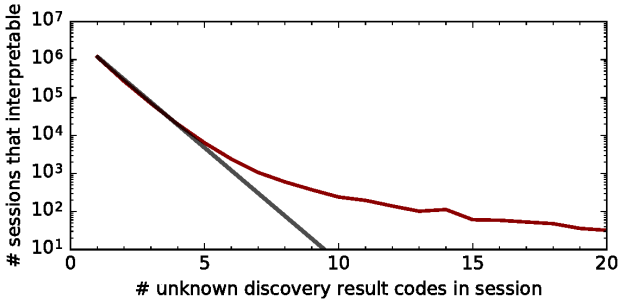



Figure 15. Discovery result code enclosed by sessions

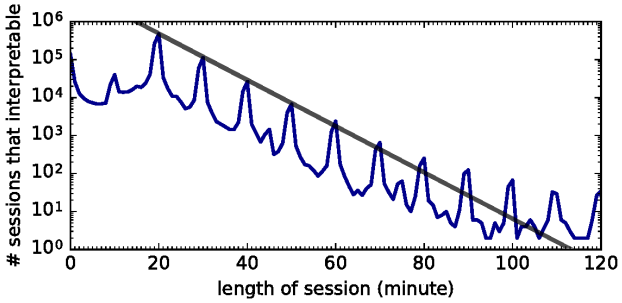



Figure 16. Length of candidate sessions

not have an exponential fit will remain FIREWALL_BLOCKS. This means that in this way we cannot prove the opposite (firewall blocks the connection). In general, we are looking for a session that begins and ends with the same network property and there are only uncertain online states between them. These sessions may be interpretable based on the begin-end enclosures. More specifically, the sessions must

- begin and end with the same NAT discovery result code
- begin and end with the same Service Set Identifier (SSID) in the case of a wifi connection
- begin and end with the same mobile operator in the case of a mobile data connection
- contain only uncertain online states
- contain a time gap between two records only in a range of 0 to 15 minutes based on the fact that the maximum time gap between two regular online records is almost 10 minutes. However, it is not very accurate because of the Android support scheduler with its inexact trigger time requirements.
- not be interrupted by trigger events that correspond to any potential change in network properties.

We show the above-defined candidate sessions in Figure 15 and Figure 16. Let us first take a look at how many uncertain discovery result codes are enclosed by these sessions in Figure 15. It is clear that the first four point seem to fit an exponential curve. Consequently, it is still open to interpretation and the rest of the points remain undefined. Next, Figure 16 shows length of the above-defined sessions. There are some peaks around every 10 minutes. These peaks correspond to the BATTERY_SCHEDULED trigger event, which is scheduled every 10 minutes and this is the most common trigger event. For example, if there is exactly one uncertain FIREWALL_BLOCKS value in the appropriate session and every taking of a measurement is triggered by this schedule event, then its length of time is around 20 minutes. Based on this example, an above-defined session that contains three unknown records lasts for 50 minutes. Accordingly, we examined the points from the first phase up to 50 minutes. Our examination revealed that it also had an exponential pattern. In contrast to the distribution in Figure 15, this distribution appears more complex, but it is still acceptable. Next we associate the two findings. More specifically, the intersection of the two sets is an above-defined session that contains less than five uncertain elements and it lasts no longer than 50 minutes. Based on this rule we can correct the network properties of 6.7% measurement records.

Next, we should mention some further minor errors associated with data collection. In a very few cases there was no network connection, but it still has some errors in the discovery result code (mainly code 0). We simply correct all of them to the no connection state (-2).

Now let us have a look at the final results of the NAT data correction in Figure 14. Records with FIREWALL_BLOCKS code are reduced to 10%, and the records with online state are expanded.

## V. LESSONS LEARNT

Based on our findings, some of the challenges encountered proved to be quite trivial, and required only some small modifications to the algorithm, while others still have to be tested with our proposed solutions.

On the client side, we have found several elements where the default approach of Android development proved insufficient, and special consideration was needed for proper data collection. We found that the timeout value of the Android application should be increased in proportion to the connectivity quality with the data collector server, while the number of retries should be reconsidered and perhaps revised with upload batches accompanied by identifiers to make duplicate detection easier.

The detection of the NAT anomalies was made significantly easier through the NetworkInfo and WifiInfo objects of the Android system. When collecting network data, we found it highly advisable to include as many attributes from these rich objects as possible - such as SSID, whether the phone is in the

 ISBN: 978-1-61208-571-5 

roaming mode and whether the network is connection metered -, since any of these could explain possible anomalies in the dataset. For example, the phone might be connected to a wifi network, but the router is not necessarily connected to the Internet; or, if it is located at a public establishment, it may redirect the requests to the establishment's login site instead of the original destination - all of which are serious problems, and they could go unnoticed without detailed information about the network.

Regarding the NAT problem, it is also advisable to reconsider storing the list of external servers in a constant array (a practice which is very common based on our experiences), because if one of those servers goes offline, it might generate huge amounts of incorrect data. A proxy which stores the server list, keeps it updated by using regular checkups, and forwards the list to the phones on request, would be a better solution here.

Also, while the deletion of previous data is a good practice to stop the application from taking up too much storage space, the 24 hour limit might be too short, since important events could get lost in that time period. The time limit for storage before deletion should be featured among the settings. Even after a delete, it is necessary to leave some trace of the deleted data - at least a log -, so the anomalies in the later, successful uploads could be interpreted.

The timestamp desynchronization between the server and the client remains perhaps the most challenging problem, with the battery based sortings providing some improvements in the dataset. One solution might be a lightweight log timestamping. In this case only a hash of the log would be sent to the server frequently (in order to minimize the mobile traffic and preserve the battery), where a reliable timestamp would be attached on the server side to this hash and saved in a permanent storage. In this way, we may define reliable milestones which are independent of the mobile side timestamps. On the mobile side, it is important to preserve the total order of the events. This could be achieved by using a simple increasing indexing procedure in the SQLite database.

We mentioned that even NAT types may be misleading, despite the quality of the connection. Once again, some of these incorrect values could be corrected by simply checking the actual state of connectivity during the upload. The NAT type in the remaining records is mostly corrected by a pattern recognition method. Hopefully, this problem may never occur again after the proposed changes have been made to handle a dynamic STUN server list.

## VI. Conclusions

As the reader can no doubt see that our approach worked well in the above-mentioned areas of data cleaning. Since the application was launched in 2013, it has been downloaded by more than 14.000 users from over 1300 different carriers and 35000 different WiFi areas, to hundreds of different device types, which is providing enormous amounts of valuable data for the analysis of NAT traces, patterns, and later on, for the simulation of the above attributes.

Compared to other crowdsourcing projects, our crowdsourcing approach was a hybrid methodology, where we provided a certain number of users with smartphones, and released the app to the Play Store for wider availability, and took more data measurements from different parts of the world. We did not reach the volume of OpenSignal or Bredbandskollen with their 100-200+ million datasets, but this hybrid solution still provided us with a much bigger amount of valuable data than a closely monitored environment like FlierMeet or SignalLab that had roughly 40 devices, a shorter collection time period, and operated in a restricted environment like a university campus or a development environment.

Lastly, though we have encountered some of the most challenging problems of the smartphone-based data-collection, our data cleaning approach successfully handled the incorrect, distorted data, and turned the dataset into a clean, organized state, which is ready for further use and processing. From this clean data, we are ready to start the graph modeling of the worldwide peer-to-peer smartphone network. In this simulated network we will be able to test the dynamics, attributes and capabilities of smartphones with real-life measurements, and later, the results of concrete peer-to-peer protocol applications. Our next major step will be the development and in-depth testing of peer-to-peer algorithms, the determination of the speed and stability of smaller applications running in this simulated environment.

## Acknowledgment

This research was supported by the Hungarian Government and the European Regional Development Fund under the grant number GINOP-2.3.2-15-2016-00037 ("Internet of Living Things").

## References

[1] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias, and D. Zeinalipour-Yazti, "Crowdsourcing with smartphones," IEEE Internet Computing, vol. 16, no. 5, 2012, pp. 36–44.

[2] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, "Crowdroid: behavior-based malware detection system for android," In Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices, October 2011, pp. 15–26, aCM.

[3] B. Guo and et al., "FlierMeet: a mobile crowdsensing system for cross-space public information reposting, tagging, and sharing," IEEE Transactions on Mobile Computing, vol. 14, no. 10, 2015, pp. 2020–2033.

[4] M. Stevens and E. D'Hondt, "Crowdsourcing of Pollution Data using Smartphones," in Workshop on Ubiquitous Crowdsourcing, held at Ubicomp '10, September 2010, pp. 1–4.

[5] A. Leemann, M. J. Kolly, R. Purves, D. Britain, and E. Glaser, "Crowdsourcing language change with smartphone applications," PloS one, vol. 11, no. 1, 2016, pp. 1–25, e0143060.

[6] T. Linder, P. Persson, A. Forsberg, J. Danielsson, and N. Carlsson, "On using crowd-sourced network measurements for performance prediction," in In Wireless On-demand Network Systems and Services (WONS), 2016 12th Annual Conference on. IEEE, January 2016, pp. 1–8.

[7] A. Overeem and et al., "Crowdsourcing urban air temperatures from smartphone battery temperatures," Geophysical Research Letters, vol. 40, no. 15, 2013, pp. 4081–4085.

[8] Á. Berta, V. Bilicki, and M. Jelasity, "Defining and understanding smartphone churn over the internet: a measurement study. In Peer-to-Peer Computing (P2P)," in 14-th IEEE International Conference on. IEEE, September 2014, pp. 1–5.

[9] E. Oliver, "The challenges in large-scale smartphone user studies," In Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement, June 2010, p. 5, aCM.

[10] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum, "LiveLab: measuring wireless networks and smartphone users in the field," ACM SIGMETRICS Performance Evaluation Review, vol. 38, no. 3, 2011, pp. 15–20.

[11] D. T. Wagner, A. Rice, and A. R. Beresford, "Device Analyzer: Large-scale mobile data collection," ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, 2014, pp. 53–56.

[12] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio, "Smartphone-based crowdsourcing for network monitoring: Opportunities, challenges, and a case study," IEEE Communications Magazine, vol. 52, no. 1, 2014, pp. 106–113.

 ISBN: 978-1-61208-571-5 

# Continue to Use Mobile Applications of Low-Cost Carriers

Edward C.S. Ku
Department of Travel Management
National Kaohsiung University of Hospitality and Tourism
Kaohsiung, Taiwan
edwardku@mail.nkuht.edu.tw

***Abstract*— The goal of this research is to determine how the firm reputation of a Low-Cost Carriers (LCCs) affects the compatibility and quality of a mobile device application. Increasing their firm reputation among passengers has become an important issue for LCCs; for an LCC to be successful, it must get rid of the impression of being cheap. In addition, LCC apps provide emotional support that is related to an individual's ability to recognize and describe his/her own or others' emotions. From the perspective of firm, LCCs need to understand the needs and analyze and behaviors of their passengers, who rely on extensive panels of instrumentation that must be checked regularly to detect updates on flights. Passengers face the same route with more than two LCC service competition. LCCs will create a process that could function as a model for their additional high-quality services.**

***Keywords- low-cost carrier; apps; continue to us;, perceived compatibility***

## I. INTRODUCTION

The emergence of Low-Cost Carriers (LCCs) has been one of the most significant developments in the aviation industry in recent years [1][2]. Additionally, transportation providers, including LCCs, now utilize mobile technologies for activities as diverse as ticketing, reservation, customer relationship, and providing real-time travel updates for their passengers [3]. However, relatively little is known about the extent to which different business strategies have adopted and integrated apps into their businesses.

LCCs emphasize e-commerce as the main channel for sales. Passengers are given access to online services and brought into a physical environment, allowing them to purchase products online and at the same time receive the products or services in a real-world environment. LCCs' apps contribute to ease of use, improve search effectiveness, and save both time and effort [3][8]. From the perspective of service-dominant (S-D) logic, LCC apps provide the customer service as the primary value, in contrast to traditional products, which offer the firm reputation as the primary value. Passengers use LCC apps to facilitate the customer–firm interactions. On the other hand, LCC business models have changed by allowing passengers to engage more interactively than ever through the use of apps. In LCC apps with an excellent navigation system, the novel, interactive information, services, and tasks offered to passengers represent either totally new experiences; these include the use of LCC apps to know about the in-flight requirements or to market existing products/services in novel ways, such as by personalizing traditional products/services.

The goal of this research is to determine how the firm reputation of an LCC affects the compatibility and quality of a mobile device application. Further, this work aims to evaluate passengers' continued app usage considering the moderating effect of service process fit. A model of the continuous use of LCC apps is presented, which, along with the hypotheses, is tested by structural equation modeling. Following the above discussion of the motivation for this study, Section 2 of this report presents the theoretical background of the research. Section 3 then provides a review of previous works, and Section discusses the research design. Finally, the research findings and conclusions are reported in Sections 5 and 6, respectively.

## II. LITERATURE REVIEW

Based on the uses and gratifications (U&G) perspective and service-dominant logic (S-D logic), we examine three major classes of persuasion determinants that are directly related to the cognitive evaluation by passengers, namely, firm reputation, perceived compatibility, and confirmation of apps.

### *A. Firm reputation*

Firm reputation refers to how customers view business services and engage in all activities, resulting in a more subjective impression of personal feelings and spread in the market stability. For LCCs, firm prestige is also based on passengers' past experiences with the airline, considering such factors as whether the LCC records the different needs of each passenger, actively cares for its passengers in a timely manner, and provides various service programs [7][11]. According to his or her relevant experiences and personal feelings, each passenger makes a comprehensive evaluation of the airline, and based on the consistent evaluation made by different passengers, the firm reputation is then created.

 ISBN: 978-1-61208-571-5 

### B. Cognitive evaluation

Cognitive value for passengers in the purchase of products, comprehensive experience, the characteristics of the product, the need to pay the time and money to do comparison will impact passenger buying behavior. Cognitive evaluation occurs most frequently and most strongly during consumption [12]. The relationship between corporate reputation and cognitive evaluation has been previously pointed out; that is, in assessing corporate reputation, it is not only the subjective understanding of business characteristics that is considered but also the impact of these features on human cognitive evaluation.

In addition, the impact of word of mouth among passengers cannot be ignored in cognitive evaluation, because third-party opinions from outside the firm are more acceptable and believable to other passengers, especially those with experience who are more persuasive than the marketing of the company, and the effect of the web. Rational evaluation can be used as an important indicator of passengers' perceptions to determine the reputation of LCCs, which affects passengers' purchase decisions [9][13]. Thus, if the company has a good reputation, it will have a considerable competitive advantage. This leads to Hypothesis 1.

*Hypothesis 1: The firm reputation of LCCs is positively associated with the cognitive evaluation of passengers.*

### C. Perceived compatibility

Compatibility refers to the extrinsic advantages of product or service consumption and usually corresponds to non–product-related attributes, especially user imagery. LCCs expect apps services to attract a large number of passengers who eventually become loyal to that mobile service; these passengers install the apps as a means to fulfill their communication and service requirements [6]. Previous research has shown that benefits exert an indirect positive influence on passengers' intention to use a mobile service by improving passenger commitment. However, reaping such benefits from apps services is possible only when the passengers are loyal to the mobile service. This leads to Hypotheses 2.

*Hypothesis 2: The perceived compatibility of apps is positively associated with the cognitive evaluation of passengers.*

### D. Confirmation capability of apps

Confirmation refers to a practical or technical advantage that users can obtain by using a specific product or service [14]. Apps enable passengers to make online reservations and to search for available trains according to departure/arrival station and date and time segment; this creates a good impression to those phases of the personal lifestyle where insignias are more valued. Similarly, Apps can assist with the distribution of information across various units of the LCC and between different levels of passengers. Passengers are more willing to use apps when they recognize the tremendous value that can be obtained from their usefulness. This leads to Hypotheses 3.

*Hypothesis 3: The confirmation capability of apps is positively associated with the cognitive evaluation of passengers.*

### E. Continued usage intention

Continued usage reflects the motivational influences that drive an individual to perform a behavior [9]. From the LCC perspective, when a passenger has experienced finding inaccurate information on an app, his or her perception of the information quality of that website will certainly hamper his or her intention to use that website again when a similar need arises. In other words, the perception of whether a particular website is qualified to help with certain tasks will influence the continuance decision. Similarly, when a passenger feels that an app is cumbersome to use or that the staff behind the LCC website are unwilling to provide sincere help, he or she may prefer to use alternative online channels. This leads to Hypothesis 4.

*Hypothesis 4: The cognitive evaluation of apps is positively associated with the continued apps usage intention of passengers.*

### F. The moderating effect of service process fit

Service process fit refers to the configuration of technologies by which service providers sense and respond to the dynamic and complex needs of passengers through information technology or online [5][10]. LCCs should analyze passengers' experiences and problems and then respond and support their needs accordingly. Apps must conform to the needs of passengers based on integrated and reliable passenger information. Therefore, we argue that congruity must exist between what individuals search for and their evaluations of the apps. Hence, we propose that passengers use apps as a support instrument when they buy tourism products and that the service process fit of an app will moderate their evaluation toward the behavioral intention use the apps. This leads to Hypothesis 5.

*Hypothesis 5: The service process fit moderates the effect of the cognitive evaluation of passengers on the continued usage of apps.*

## III. METHODOLOGY

Stratified sampling was used in data collection, which was done through a field survey; we confirmed that the selected passengers had downloaded and used the

 ISBN: 978-1-61208-571-5 

apps which they take LCC in advance, and then we asked them if they were willing to fill in the questionnaire; passengers that did not have LCC apps usage experience were excluded from the study.

We confirmed that the selected passengers had downloaded and used the apps which they take LCC in advance, and then we asked them if they were willing to fill in the questionnaire; passengers that did not have LCC apps usage experience were excluded from the study. A total of 1,500 passengers were invited based on the sampling strategy used in the airports; 888 of them completed the questionnaires (59.2% return rate). The questionnaires were developed based on a literature review on related topics.

## IV. ANALYSIS AND RESULTS

The measurement items were taken from the literature, and the constructs in the study were measured with a multi-item scale; the internal consistency (Cronbach's alpha) of the construct was greater than 0.9 [15]. The LISREL 8.50 software was used in the analysis. An AVE estimate of 0.50 or higher indicates acceptable validity for a construct measure. The results show that the square root of all AVE estimates for each construct is greater than the inter-construct correlations, as shown in Table I, thus, discriminant validity is supported. The overall goodness of fit was assessed in terms of eight common model fit measures: GFI, 0.92; AGFI, 0.92; RMR, 0.051; RMSEA, 0.058; NFI, 0.92; PNFI, 0.72, CFI, 0.91; and PGFI, 0.66. The $\chi 2$ degrees of freedom ratio was 1.69. Overall, the data indicate a good fit for our hypothesized model, and the results provide meaningful support for the research hypotheses. Among the five hypotheses, five are fully supported.

TABLE I MODEL ESTIMATION

| | FR | PCO | Con | CE | SPF | CUI | AVE |
|---|---|---|---|---|---|---|---|
| FR | 0.714 | | | | | | 0.51 |
| PCO | 0.487 | 0.843 | | | | | 0.71 |
| Con | 0.251 | 0.307 | 0.943 | | | | 0.89 |
| CE | 0.456 | 0.407 | 0.290 | 0.781 | | | 0.61 |
| SPF | 0.442 | 0.480 | 0.295 | 0.471 | 0.938 | | 0.88 |
| CUI | 0.451 | 0.231 | 0.326 | 0.364 | 0.441 | 0.872 | 0.76 |

The main diagonal shows the square of the AVE (averaged variance extracted).
Significance at p <0.05 level is shown in hold.
FR stand for firm reputation, PCO for perceived compatibility, Con for confirmation, CE for cognitive evaluation, SPF for service process fit, CUI for continued usage intention.

The results indicated a good fit, overall, and the data indicate a favorable fit for our hypothesized model. Table II shows the structural model shows all hypotheses are accepted.

TABLE II HYPOTHESES AND RESULTS

| | Hypothesis | T-value | Results |
|---|---|---|---|
| H1 | Firm Reputation -> Cognitive Evaluation | 4.5** | Supported |
| H2 | Perceived Compatibility -> Cognitive Evaluation | 11.75** | Supported |
| H3 | Confirmation -> Cognitive Evaluation | 6.92** | Supported |
| H4 | Cognitive Evaluation -> Continued Usage Intention | 15.63*** | Supported |
| H5 | Cognitive Evaluation* Service Process Fit -> Continued Usage Intention | 3.93* | Supported |

$p<.05$, $p<.01$, $p<.001$

## V. CONCLUSIONS AND IMPLICATIONS

The use of structural equation modeling to test the theoretical model of continuous usage could lead to a better understanding of the nature and determinants of choices and decisions related to LCCs. The present study focused on continuous LCC apps usage by applying a multidimensional measure of factors that influence continuous usage, with the said measure being both intuitively appealing and reliable. The analysis of the measurement model indicates that the proposed metrics have an acceptable degree of validity and reliability.

First, increasing their firm reputation among passengers has become an important issue for LCCs; for an LCC to be successful, it must get rid of the impression of being cheap. As the LCC increases its service quality, for example, by improving the punctuality rate of flight, the mechanical maintenance, and the image formed by word of mouth, its reputation will become satisfactory to passengers; accordingly, the LCC will improve its business performance, market share, shareholder returns, and passenger service.

Most passengers believe that LCCs focus on cost-oriented segmentation; however, given the growth of the LCC market, LCC need to boost their reputation and gain passengers' recognition. As the LCC increases its service quality, for example, by improving the punctuality rate of flight, the mechanical maintenance, and the image formed by word of mouth, its reputation will become satisfactory to passengers

Second, the interface design of the apps provide a personalized presentation that reflects the preferences of the passengers. Moreover, emotion-related compatibility can be characterized as a combination of psychological tool demands and the tasks to which passengers can decide for themselves how to do their usage behavior. Thus, LCCs should enhance the design of their apps to attract more passengers to use them.

LCC apps provide emotional support that is related to an individual's ability to recognize and describe his/her

 ISBN: 978-1-61208-571-5 

own or others' emotions. Moreover, emotion-related compatibility can be characterized as a combination of psychological tool demands and the tasks to which passengers can decide for themselves how to do their usage behavior. Thus, LCCs should enhance the design of their apps to attract more passengers to use them.

Third, the confirmation capability of apps that work in dynamic environments continually monitor systems; LCCs have been described as producing a broad range of service quality improvements. LCCs need to analyze the needs and behaviors of their passengers, who rely on extensive panels of instrumentation that must be checked regularly to detect updates on flights. Similarly, keeping an open channel for communication with passengers, providing information during periods of apps usage, and encouraging two-way communication lead to higher passenger satisfaction. Thus, the provision of service processes in the apps makes passengers experience the value of apps usage.

LCCs need to keep an open channel for communication with passengers, providing information during periods of apps usage, and encouraging two-way communication lead to higher passenger satisfaction. Thus, the provision of service processes in the apps makes passengers experience the value of apps usage.

Finally, service process fit in operations is necessary to make apps functions consistent with the service-oriented work process. Similarly, LCCs can create a process that could function as a model for their additional high-quality services. For example, the system can automatically generate the two-dimensional bar code for baggage within the apps, and the apps can record the baggage weight information for passengers, which will help passengers in the fare on the expenditure judgement. In addition, LCCs can estimate the overbooking by applying system analysis, as well as develop a unique overbooking and transit service strategy.

## REFERENCES


[1] T. Buaphiban, and D. Truong, "Evaluation of passengers' buying behaviors toward low cost carriers in Southeast Asia." Journal of Air Transport Management vol.59, 2017, pp.124-133.

[2] L.Y. Chang, and S.C. Hung, Adoption and loyalty toward low cost carriers: "The case of Taipei–Singapore passengers." Transportation Research Part E: Logistics and Transportation Review, vol.50, 2013, pp. 29-36.

[3] S. C. Chiu, C. H. Liu, J. H. Tu, "The influence of tourists' expectations on purchase intention: Linking marketing strategy for low-cost airlines." Journal of Air Transport Management vol.53, 2016, pp.226-234.

[4] A. Y. L. Chong, "Understanding mobile commerce continuance intentions: an empirical analysis of Chinese consumers." Journal of Computer Information Systems vol.53, 2016, pp.22-30.

[5] Y. W. Fan, and E. Ku, "Customer focus, service process fit and customer relationship management profitability: the effect of knowledge sharing." The Service Industries Journal vol.30, 2010, pp.203-223.

[6] X. Fu, Z. Lei, K. Wang, and J. Yan, "Low cost carrier competition and route entry in an emerging but regulated aviation market–The case of China. Transportation Research Part A: Policy and Practice vol.79, 2015, pp.3-16.

[7] H. Han, "Effects of in-flight ambience and space/function on air travelers' decision to select a low-cost airline." Tourism management vol.37, 2013, pp.125-135.

[8] C. I. Ho, M. H. Lin, H. M. Chen, "Web users' behavioural patterns of tourism information search: From online to offline." Tourism Management vol.33, 2012, pp.1468-1482.

[9] H. Hoehle, X. Zhang, and V. Venkatesh, "An espoused cultural perspective to understand continued intention to use mobile applications: a four-country study of mobile social media application usability." European Journal of Information Systems vol.24, 2015, pp.337-359.

[10] E. Ku, "Functional integration and systems implementation of customer relationship management in hotel industry: A multilevel analysis." International Journal of Information Technology & Decision Making vol.13, 2014. Pp.175-196..

[11] K. Seo, J. Moon, and S. Lee, "Synergy of corporate social responsibility and service quality for airlines: The moderating role of carrier type." Journal of Air Transport Management vol.47, 2015, pp.126-134.

[12] I. Vlachos, and Z. Lin, "Drivers of airline loyalty: Evidence from the business travelers in China." Transportation Research Part E: Logistics and Transportation Review vol.71, 2014, pp.1-17.

[13] Y. Wang, K. K. F.So, and B. A.Sparks, "What technology-enabled ServiceS do air travelerS value? inveStigating the role of technology readiness." Journal of Hospitality & Tourism Research, 2015 1096348014538050.

[14] T. Zhou, "An empirical examination of users' post-adoption behaviour of mobile services." Behaviour & Information Technology vol.30, 2011, pp.241-250.

[15] R. Brown, "A cautionary note on the use of LISREL'S automatic start values in confirmatory factor analysis studies. Applied psychological measurement," vol.10, 1986, pp.239-245.


.

 ISBN: 978-1-61208-571-5 