

Nguyen-Thinh Le  
Tien Van Do  
Ngoc Thanh Nguyen  
Hoai An Le Thi *Editors*

# Advanced Computational Methods for Knowledge Engineering

Proceedings of the 5th International  
Conference on Computer Science,  
Applied Mathematics and  
Applications, ICCSAMA 2017

# **Advances in Intelligent Systems and Computing**

Volume 629

## **Series editor**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland  
e-mail: [kacprzyk@ibspan.waw.pl](mailto:kacprzyk@ibspan.waw.pl)

### *About this Series*

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within “Advances in Intelligent Systems and Computing” are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

### *Advisory Board*

#### Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

e-mail: [nikhil@isical.ac.in](mailto:nikhil@isical.ac.in)

#### Members

Rafael Bello Perez, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba

e-mail: [rbellop@uclv.edu.cu](mailto:rbellop@uclv.edu.cu)

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

e-mail: [escorchado@usal.es](mailto:escorchado@usal.es)

Hani Hagrass, University of Essex, Colchester, UK

e-mail: [hani@essex.ac.uk](mailto:hani@essex.ac.uk)

László T. Kóczy, Széchenyi István University, Győr, Hungary

e-mail: [koczy@sze.hu](mailto:koczy@sze.hu)

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan

e-mail: [ctlin@mail.nctu.edu.tw](mailto:ctlin@mail.nctu.edu.tw)

Jie Lu, University of Technology, Sydney, Australia

e-mail: [Jie.Lu@uts.edu.au](mailto:Jie.Lu@uts.edu.au)

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico

e-mail: [epmelin@hafsamx.org](mailto:epmelin@hafsamx.org)

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil

e-mail: [nadia@eng.uerj.br](mailto:nadia@eng.uerj.br)

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland

e-mail: [Ngoc-Thanh.Nguyen@pwr.edu.pl](mailto:Ngoc-Thanh.Nguyen@pwr.edu.pl)

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong

e-mail: [jwang@mae.cuhk.edu.hk](mailto:jwang@mae.cuhk.edu.hk)

More information about this series at <http://www.springer.com/series/11156>

Nguyen-Thanh Le · Tien Van Do  
Ngoc Thanh Nguyen · Hoai An Le Thi  
Editors

# Advanced Computational Methods for Knowledge Engineering

Proceedings of the 5th International  
Conference on Computer Science,  
Applied Mathematics and  
Applications, ICCSAMA 2017

 Springer

*Editors*

Nguyen-Thinh Le  
Department of Informatics  
Humboldt-Universität zu Berlin  
Berlin  
Germany

Ngoc Thanh Nguyen  
Department of Information Systems  
Wrocław University of Science  
and Technology  
Wrocław  
Poland

Tien Van Do  
Department of Networked Systems  
and Services  
Budapest University of Technology  
and Economics  
Budapest  
Hungary

Hoai An Le Thi  
Theoretical and Applied Computer Science  
Laboratory  
University of Lorraine  
Metz  
France

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-3-319-61910-1

ISBN 978-3-319-61911-8 (eBook)

DOI 10.1007/978-3-319-61911-8

Library of Congress Control Number: 2017943843

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This volume contains papers presented at the 5th *International Conference on Computer Science, Applied Mathematics and Applications* (ICCSAMA 2017) held on June 30 and July 1, 2017, in Berlin, Germany. The conference is co-organized by research group “Computer Science Education/Computer Science and Society” at Department of Informatics, Humboldt-Universität zu Berlin (Germany), Analysis, Design and Development of ICT systems (AddICT) Laboratory, Budapest University of Technology and Economics (Hungary), Division of Knowledge Management Systems, Wrocław University of Science and Technology (Poland), and Theoretical and Applied Computer Science Laboratory, University of Lorraine (France) in cooperation with IEEE SMC Technical Committee on Computational Collective Intelligence.

The aim of ICCSAMA 2017 is to bring together leading academic scientists, researchers, and scholars to discuss and share their newest results in the fields of computer science, applied mathematics, and their applications. After the peer review process, 19 papers by authors from Austria, Chile, Ecuador, France, Germany, Hungary, Italy, India, Japan, Poland, Spain, Thailand, UK, and Vietnam have been selected for including in this proceedings.

In particular, ICCSAMA 2017 hosts the special session “Human Computer Communication and Intelligent Applications,” which has been proposed and organized by Thepchai Supnithi (National Electronics and Computer Technology Center, Thailand), Thanaruk Theeramunkong (Thammasat University, Thailand), Tomoko Kojiri (Kansai University, Japan), Mahasak Ketcham (King Mongkut’s University of Technology North Bangkok, Thailand), and Christoph Benzmüller (Freie Universität Berlin, Germany). In addition, the conference program is enriched by two keynotes that are given by Prof. Hiroshi Tsuji, President of Osaka Prefecture University (Japan), and Prof. Tomoko Kojiri, Kansai University (Japan).

The clear message of the proceedings is that the potentials of computational methods for knowledge engineering and optimization algorithms are to be exploited, and this is an opportunity and a challenge for researchers. It is observed that the ICCSAMA 2013–2016 clearly generated a significant amount of interaction between members of both communities on computer science and applied

mathematics. The intensive discussions have seeded future exciting development at the interface between computational methods, optimization, and engineering.

The works included in these proceedings would be useful for researchers, and Ph.D. and graduate students in optimization theory and knowledge engineering fields. It is the hope of the editors that readers can find many inspiring ideas and new research directions for their research. Many such challenges are suggested by particular approaches and models presented in the proceedings.

We would like to thank all authors, who contributed to the success of the conference and to this book. Special thanks go to the members of the Steering and Program Committees for their contributions to keeping the high quality of the selected papers. Cordial thanks are due to the Organizing Committee members for their efforts and the organizational work. We acknowledge Humboldt-Universität zu Berlin for hosting the conference ICCSAMA 2017. Finally, we cordially thank Springer for supports and publishing this volume.

We hope that ICCSAMA 2017 significantly contributes to the fulfillment of the academic excellence and leads to greater success of ICCSAMA events in the future.

July 2017

Nguyen-Thinh Le  
Tien Van Do  
Hoai An Le Thi  
Ngoc Thanh Nguyen

# ICCSAMA 2017 Organization

## General Chair

Nguyen-Thinh Le                      Humboldt-Universität zu Berlin, Germany

## General Co-chair

Tien Van Do                              Budapest University of Technology and  
Economics, Hungary

## Program Chairs

Tien Van Do                              Budapest University of Technology and  
Economics, Hungary  
Hoai An Le Thi                            University of Lorraine, France  
Niels Pinkwart                            Humboldt-Universität zu Berlin, Germany

## Organizing Committee

Nam Hoai Do                              Budapest University of Technology and  
Economics, Hungary  
Nguyen-Thinh Le                            Humboldt-Universität zu Berlin, Germany

## Steering Committee

Tien Van Do                              Budapest University of Technology and  
Economics, Hungary



Dosam Hwang	Yeungnam University, Korea
Hoai An Le Thi	University of Lorraine, France (Co-chair)
Ngoc Thanh Nguyen	Wroclaw University of Science and Technology, Poland (Co-chair)
Dinh Tao Pham	INSA Rouen, France
Duc Truong Pham	University of Birmingham, UK
Hoang Pham	Rutgers, The State University of New Jersey, USA
Phuoc Tran-Gia	University of Würzburg, Germany

## Program Committee

Lars Brauchbach	University of Hamburg, Germany
Alain Bui	Université de Versailles-St-Quentin-en-Yvelines, France
Tien Van Do	Budapest University of Technology and Economics, Hungary
Nam Hoai Do	Budapest University of Technology and Economics, Hungary
Fabio Gasparetti	ROMA TRE University, Italy
Sebastian Gross	Humboldt-Universität zu Berlin, Germany
Christian Guetl	Graz University of Technology, Austria
Yuki Hayashi	Osaka Prefecture University, Japan
Sharon I-Han Hsiao	Arizona State University, USA
Dosam Hwang	Yeungnam University, Korea
Tomoko Kojiri	Kansai University, Japan
Amruth N. Kumar	Ramapo College of New Jersey, USA
Ki-Ryong Kwon	Pukyong National University, Korea
Hoai Minh Le	University of Lorraine, France
Hoang-Son Le	VNU University of Science, Vietnam National University, Vietnam
Nguyen-Thinh Le	Humboldt-Universität zu Berlin, Germany
Duc-Hau Le	Water Resources University, Vietnam
Hoai An Le Thi	University of Lorraine, France
Andreas Lingnau	Humboldt-Universität zu Berlin, Germany
Jon Mason	Charles Darwin University, Australia
Noboru Matsuda	Texas A&M University, USA
Ngoc Thanh Nguyen	Wroclaw University of Science and Technology, Poland
Thanh Binh Nguyen	IIASA, Austria
Loan T.T. Nguyen	Nguyen Tat Thanh University, Vietnam
Van Thoai Nguyen	Universität Trier, Germany
Asako Ohno	Osaka Sangyo University, Japan
Niels Pinkwart	Humboldt-Universität zu Berlin, Germany
Duc Truong Pham	University of Birmingham, UK

Dinh Tao Pham	INSA Rouen, France
Hoang Pham	Rutgers, The State University of New Jersey, USA
Alexander Pokahr	University of Hamburg, Germany
Hoang Quang	Hue University, Vietnam
Martina Rau	University of Wisconsin, Madison, USA
Ma. Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Yanjie Song	The Education University of Hong Kong, Hong Kong
Thepchai Supnithi	National Electronics and Computer Technology Center, Thailand
Ha Quang Thuy	Vietnam National University, Vietnam
Simon Tjoa	St. Poelten University of Applied Sciences, Austria
Phuoc Tran-Gia	University of Würzburg, Germany
Bay Vo	Ho Chi Minh City University of Technology, Vietnam
Florian Wamser	University of Würzburg, Germany
Longkai Wu	Nanyang Technological University, Singapore
Thomas Zinner	University of Würzburg, Germany

# **Keynotes**

# Forty Years Experience on R&D and Education of Systems Sciences

Hiroshi Tsuji

Osaka Prefecture University, Osaka, Japan

**Abstract.** In order to appeal the importance of system thinking, I will review three cases and introduce the education in Osaka Prefecture University. The system thinking in my context includes long-term view, wide-ranging view, and upside-down view. Even if the techniques do not seem to be available for practical use at first, it often becomes available once the TPO (Time, Place and Occasion) change. If one has only a hammer, he can only look for nails to solve his problem. However, if he has a variety of tools, he has also chance to select a right tool to solve his problem. Case 1 relates to the regression model. When I joined Hitachi, Ltd., I encountered the problem for resource allocation. Although our approach based on regression model did not work then, we apply this technique to risk analysis on offshore outsourcing twenty five years later. Case 2 relates to version space proposed in the middle of 1980s. When I learned this technique at CMU, I cannot imagine which domain the technique can be applied to. However, as I learned it, I could apply this idea to develop an expert system for program tuning. Case 3 relates to interpretive structural modeling which I learned while I was a graduate student. Now, we have chance to apply this technique to a big data on learning log. Finally, I will introduce some photographs which are metaphor of system thinking.

# How We Grasp “Causal Relation” in Historical Learning?

Tomoko Kojiri

Faculty of Engineering Science, Kansai University, Suita, Japan

**Abstract.** The concept map is a powerful externalization tool for representing the relationships between concepts or ideas. This type of map is also used in many educational settings to make students deeply consider the target knowledge. One of the effective uses of applying a concept map is to arrange the historical events with the causal relationship in the historical learning. The created map makes students aware of the flow of history and also important factors that cause significant events. However, some students find difficulty in grasping “causal relation” between historical events and are not able to create the concept map. Our research group has been pursuing the effective way of grasping “causal relation” in history and has developed two support systems for creating valid concept map of historical events. In this talk, we introduce our two systems and show the effects of these systems through the trial use by university students and junior high school students. In addition, we show the result of introducing one of these systems into the classroom of the junior high school.

# Contents

<b>Stochastic DCA for Sparse Multiclass Logistic Regression</b> . . . . .	1
Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, and Bach Tran	
<b>Reformulation of the Quadratic Multidimensional Knapsack Problem as Copositive/Completely Positive Programs</b> . . . . .	13
D.V. Nguyen	
<b>DC Programming and DCA for Enhancing Physical Layer Security in Amplify-and-Forward Relay Beamforming Networks Based on the SNR Approach</b> . . . . .	23
Nguyen Nhu Tuan and Dang Vu Son	
<b>A Cash-Flow-Based Optimization Model for Corporate Cash Management: A Monte-Carlo Simulation Approach</b> . . . . .	34
Linus Krumrey, Mahdi Moeini, and Oliver Wendt	
<b>Demand Side Management: A Case for Disruptive Behaviour</b> . . . . .	47
Dina Subkhankulova, Artem Baklanov, and David McCollum	
<b>Enhancing Reduced Order Model Predictive Control for Autonomous Underwater Vehicle</b> . . . . .	60
Prashant Bhopale, Pratik Bajaria, Navdeep Singh, and Faruk Kazi	
<b>Comparison of Feedback Strategies for Supporting Programming Learning in Integrated Development Environments (IDEs)</b> . . . . .	72
Jarno Coenen, Sebastian Gross, and Niels Pinkwart	
<b>Using Online Synchronous Interschool Tournaments to Boost Student Engagement and Learning in Hands-On Physics Lessons</b> . . . . .	84
Roberto Araya, Carlos Aguirre, Patricio Calfucura, and Paulina Jaure	
<b>Story-Based Multimedia Analysis Using Social Network Technique</b> . . . .	95
Quang Hai Bang Tran, Thi Hai Binh Nguyen, Phong Nha Tran, Thi Thanh Nga Tran, and Quang Dieu Tran	

**Plot-Creation Support System for Writing Novels** . . . . . 107  
 Atsushi Ashida and Tomoko Kojiri

**An Improved Algorithm for Mining Top-k Association Rules**. . . . . 117  
 Linh T.T. Nguyen, Loan T.T. Nguyen, and Bay Vo

**A Deep Architecture for Sentiment Analysis of News Articles**. . . . . 129  
 Dinh Nguyen, Khuong Vo, Dang Pham, Mao Nguyen, and Tho Quan

**Sentiment Polarity Detection in Social Networks:**

**An Approach for Asthma Disease Management** . . . . . 141  
 Harry Luna-Aveiga, José Medina-Moreira, Katty Lagos-Ortiz,  
 Oscar Apolinario, Mario Andrés Paredes-Valverde,  
 María del Pilar Salas-Zárate, and Rafael Valencia-García

**An Overview of Information Discovery Using Latent  
 Semantic Indexing**. . . . . 153  
 Roger Bradford

**Similarity Measures for Music Information Retrieval** . . . . . 165  
 Michele Della Ventura

**An Early-Biologisation Process to Improve the Acceptance  
 of Biomimetics in Organizations** . . . . . 175  
 Nguyen-Truong Le, Joachim Warschat, and Tobias Farrenkopf

**Bidirectional Deep Learning of Context Representation for Joint  
 Word Segmentation and POS Tagging** . . . . . 184  
 Prachya Boonkwan and Thepchai Supnithi

**A Model for a Computing Cluster  
 with Two Asynchronous Servers** . . . . . 197  
 Hai T. Nguyen and T.V. Do

**A Review of Technologies for Conversational Systems** . . . . . 212  
 Julia Masche and Nguyen-Thinh Le

**Author Index**. . . . . 227

# Stochastic DCA for Sparse Multiclass Logistic Regression

Hoai An Le Thi, Hoai Minh Le<sup>(✉)</sup>, Duy Nhat Phan, and Bach Tran

Laboratory of Theoretical and Applied Computer Science - LITA EA 3097,  
University of Lorraine, Ile du Sauley, 57045 Metz, France  
{hoai-an.le-thi,minh.le,duy-nhat.phan,bach.tran}@univ-lorraine.fr  
<http://www.lita.univ-lorraine.fr/~lethi/index.php/en/>

**Abstract.** In this paper, we deal with the multiclass logistic regression problem, one of the most popular supervised classification method. We aim at developing an efficient method to solve this problem for large-scale datasets, i.e. large number of features and large number of instances. To deal with a large number of features, we consider feature selection method evolving the  $l_{\infty,0}$  regularization. The resulting optimization problem is non-convex for which we develop a stochastic version of DCA (Difference of Convex functions Algorithm) to solve. This approach is suitable to handle datasets with very large number of instances. Numerical experiments on several benchmark datasets and synthetic datasets illustrate the efficiency of our algorithm and its superiority over well-known methods, with respect to classification accuracy, sparsity of solution as well as running time.

**Keywords:** DC programming · Stochastic DCA · Sparse multiclass logistic regression

## 1 Introduction

In machine learning, supervised learning consists in building a predictor function, based on a labeled training data, which can identify the label of new instances with highest possible accuracy. Logistic regression, introduced by Cox in 1958 [2], is undoubtedly one of the most popular supervised learning methods. Logistic regression can be seen as an extension of linear regression where the dependent variable is categorical. Instead of estimating the outcome by a continuous value like linear regression, logistic regression tries to predict the probability that an instance belongs to a category. Logistic regression has been successfully applied in various real-life problems such as cancer detection [5], risk analysis [19], social science [6], etc.

In this paper, we deal with the multiclass logistic regression problem where the dependent variable has more than two outcome categories. We aim at developing an efficient method for solving the multiclass logistic regression problem to deal with data that has not only a large number of features but also large number of instances.



The multiclass logistic regression problem can be described as follows. Let  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  be a training set with observation vectors  $x_i \in \mathbb{R}^d$  and labels  $y_i \in \{1, \dots, Q\}$  where  $Q$  is the number of classes. Let  $W$  be the  $d \times Q$  matrix whose columns are  $W_{:,1}, \dots, W_{:,Q}$  and  $b = (b_1, \dots, b_Q) \in \mathbb{R}^Q$ . The couple  $(W_{:,i}, b_i)$  forms the hyperplane  $f_i := W_{:,i}^T x + b_i$  that separates the class  $i$  from the other classes.

In the multiclass logistic regression problem, the conditional probability  $p(Y = y|X = x)$  that an instance  $x$  belongs to a class  $y$  is defined as

$$p(Y = y|X = x) = \frac{\exp(b_y + W_{:,y}^T x)}{\sum_{k=1}^Q \exp(b_k + W_{:,k}^T x)}. \quad (1)$$

Then, the predicted class  $y^*$  of a new observation  $x^*$  is computed by

$$y^* = \arg \max_{k=1..Q} p(Y = k|X = x^*) = \arg \max_{k=1..Q} (b_k + W_{:,k}^T x^*). \quad (2)$$

We aim to find  $(W, b)$  for which the total probability of the training observations  $x_i$  belonging to its correct classes  $y_i$  is maximized. A natural way to estimate  $(W, b)$  is to maximize the log-likelihood function which is defined by

$$\mathcal{L}(W, b) := -\frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i, W, b), \quad (3)$$

where  $\ell(x_i, y_i, W, b) = \log \left( \sum_{k=1}^Q \exp(b_k + W_{:,k}^T x_i) \right) - b_{y_i} - W_{:,y_i}^T x_i$ . Hence, the multiclass logistic regression problem can be written as

$$\max_{W, b} \mathcal{L}(W, b) \Leftrightarrow \min_{W, b} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i, W, b) \right\}. \quad (4)$$

Clearly, the above optimization problem is convex and can be solved by a convex optimization method. However, how to efficiently solve the problem, especially in real-world applications where the datasets are large-scale, remains an open question.

On one hand, in many applications such as information retrieval, face recognition, microarray analysis, etc., datasets contain a very large number of features. In such datasets, we often encounter the problem of redundant features (information already presented by other features) and irrelevant features (features that do not contain useful information). Feature selection methods that try to select only useful features for the considered task, is a popular and efficient way to deal with this problem. A natural way to deal with feature selection problem is to formulate it as a minimization of the  $\ell_0$ -norm (or  $\|\cdot\|_0$ ) where the  $\ell_0$ -norm of  $x \in \mathbb{R}^d$  is defined by the number of non-zero components of  $x$ , namely,  $\|x\|_0 := |\{i = 1, \dots, n : x_i \neq 0\}|$ . It is well-known that the problem

of minimizing  $\ell_0$ -norm is NP-hard [1] due to the discontinuity of  $\ell_0$ -norm. The minimization of  $\ell_0$ -norm has been extensively studied on both theoretical and practical aspects for individual variable selection in many practical problems. An extensive overview of existing approaches for the minimization of  $\ell_0$ -norm can be found in [12].

In our problem, as mentioned above, the class  $i$  ( $i = 1, \dots, Q$ ) is separated from the other classes by the hyperplane  $f_i := W_{:,i}^T x + b_i$ . If the weight  $W_{j,i}$  is equal to zero then we can say that the feature  $j$  ( $j = 1, \dots, d$ ) is not necessary to separate class  $i$  from the other classes. Hence, the feature  $j$  is to be removed if and only if it is not necessary for any separator  $f_i$  ( $i = 1, \dots, Q$ ), i.e. all components in the row  $j$  of  $W$  are zero. Recently, the mixed-norm regularizations, e.g.  $\ell_{2,1}$ -norm [23] (group lasso),  $\ell_{2,0}$ -norm [20, 21] (group MCP and group SCAD) or  $\ell_{\infty,1}$ -norm [18], have been developed for this purpose. In this work, we consider the  $\ell_{\infty,0}$  regularization. This regularization term forces the weights of the same features across all classes to become zero. Denote by  $W_{j,:}$  the  $j$ -th row of the matrix  $W$ . The  $\ell_{\infty,0}$ -norm of  $W$ , i.e., the number of non-zero rows of  $W$ , is defined by

$$\|W\|_{\infty,0} = |\{j \in \{1, \dots, d\} : \|W_{j,:}\|_{\infty} \neq 0\}|.$$

Thus, the  $\ell_{\infty,0}$ -norm multiclass logistic regression problem is formulated as follows

$$\min_{W,b} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i, W, b) + \lambda \|W\|_{\infty,0} \right\}, \quad (5)$$

where  $\lambda$  is a regularization parameter or trade-off parameter that makes the trade-off between classification accuracy term and the sparsity term. Since  $\ell_{\infty,0}$ -norm is non-convex, the above optimization problem is non-convex.

Several works have been developed to solve the problem of mixed-norm regularization  $\ell_{\infty,0}$ . These works have used the convex approximation approach which replaces  $\ell_{\infty,0}$ -norm by the  $\ell_{\infty,1}$ -norm. This approach was widely used for selecting groups of variables in linear regression [18], in multi-task learning [17], multiclass support vector machine [24], etc. The sparse multiclass logistic regression using  $\ell_{\infty,1}$ -norm can be written as follows

$$\min_{W,b} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i, W, b) + \lambda \|W\|_{\infty,1} \right\}. \quad (6)$$

Since the  $\ell_{\infty,1}$ -norm is convex, the resulting problem (6) remains convex.

*Paper's Contribution:* In our work, we consider a non-convex approximation of the  $\ell_{\infty,0}$ -norm. Non-convex approximations will lead to a non-convex optimization problem which is difficult to solve. However, non-convex approximations have been proved to be more efficient than convex approximation [12]. Among the well-known non-convex approximations, we adopt the piecewise exponential approximation function. This approximation has been used in several problems. In Le Thi et al. (2008) [8], the piecewise exponential function was reformulated as

a DC function and an efficient DCA was developed for solving the problem of feature selection in SVM. The same DC decomposition was also successfully applied to feature selection in semi-supervised support vector machines [7], sparse multiclass support vector machines [9], sparse signal recovery [10], sparse scoring problem [13]. Motivated by the efficiency of the DC decomposition proposed in [8], we investigate a similar DC decomposition for the sparse multiclass logistic regression problem (5). By using this DC decomposition, the approximate problem of (5) is a DC problem. Thus, we will develop efficient algorithm based on DC programming and DCA to solve it.

On the other hand, when the number of instances  $n$  is huge, the problem (5) becomes more challenging in machine learning due to the per-iteration cost might be expensive. In the last few years, stochastic optimization techniques have been proved to be useful in machine learning for solving large-scale problems. Stochastic programming is suitable for problem (5) since it can exploit the advantage of the sum structure of (5). However, as mentioned above, the regularization term  $\ell_{\infty,0}$  makes the problem (5) non-smooth non-convex. In this paper, we propose a stochastic version of DCA which considers only a small subset of components  $\ell(x_i, y_i, W, b)$  at each iteration instead of using all  $n$  components for the calculation of  $W$  and  $b$ . Furthermore, we exploit the special structure of the problem to propose an efficient DC decomposition for which the corresponding stochastic DCA scheme is very inexpensive.

We perform an empirical comparison with two standard methods on very-large synthetic and real-world datasets, and show that our proposed algorithm is efficient in feature selection and classification as well as running time.

The remainder of the paper is organized as follows. The outline of DC programming and DCA as well as Stochastic DCA are presented in Sect. 2.1. Solution method based on Stochastic DCA is developed in Sect. 2.2. Computational experiments are reported in Sect. 3 and finally Sect. 4 concludes the paper.

## 2 Solution Method via Stochastic DCA

### 2.1 Outline of DC Programming, DCA and Stochastic DCA

DC programming and DCA constitute the backbone of smooth/non-smooth non-convex programming and global optimization [11, 14, 15]. They address the problem of minimizing a DC function on the whole space  $\mathbb{R}^n$  or on a closed convex set  $\Omega \subset \mathbb{R}^n$ . Generally speaking, a standard DC program takes the form:

$$\alpha = \inf\{F(x) := G(x) - H(x) \mid x \in \mathbb{R}^n\} \quad (P_{dc}),$$

where  $G, H$  are lower semi-continuous proper convex functions on  $\mathbb{R}^n$ . Such a function  $F$  is called a DC function, and  $G - H$  is a DC decomposition of  $F$  while  $G$  and  $H$  are the DC components of  $F$ . A DC program with convex constraint  $x \in \Omega$  can be equivalently expressed as an unconstrained DC program by adding the indicator function  $\chi_{\Omega}$  ( $\chi_{\Omega}(x) = 0$  if  $x \in \Omega$  and  $+\infty$  otherwise) to the first DC component  $G$ .

Starting from an initial point  $x^0$ , the DCA consists in constructing two sequences  $\{x^l\}$  and  $\{y^l\}$  such that  $y^l \in \partial H(x^l)$  and  $x^{l+1} \in \arg \min_{x \in \mathbb{R}^n} \{G(x) - \langle y^l, x \rangle\}$ .

Consider now a DC program of the form

$$\min \left\{ \frac{1}{n} \sum_{i=1}^n f_i(x) \mid x \in \mathbb{R}^n \right\} \quad (7)$$

where  $f_i = g_i - h_i$  is a DC function. When  $n$  is large, the problem (7) becomes more challenging in machine learning. The per-iteration cost of DCA might be more expensive because it uses all the DC functions  $f_i$ . A stochastic version of DCA and its convergence properties for some special cases were developed in [16]. The stochastic DCA for solving the problem (7) is described as follows. Starting from an initial point  $x^0$ , at each iteration  $l$ , we randomly choose one index  $i_l \in \{1, \dots, n\}$ , and compute  $y^l \in \partial h_{i_l}(x^l)$ , and then compute  $x^{l+1}$  by solving the following convex problem

$$\min \left\{ \frac{1}{l+1} \sum_{h=0}^l [g_{i_h}(x) - \langle y^h, x \rangle] \right\}.$$

## 2.2 Stochastic DCA for Solving the Sparse Multiclass Logistic Regression Problem

By using the piecewise exponential function [8], the corresponding approximate problem of (5) takes the form:

$$\min_{W,b} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i, W, b) + \lambda P(W) \right\}, \quad (8)$$

where  $P(W) = \sum_{j=1}^d \eta_\alpha(\|W_{j,:}\|_\infty)$  with  $\eta_\alpha(t) = 1 - \exp(-\alpha|t|)$ . The problem (8) can be rewritten as the following problem.

$$\min_{W,b} \left\{ f(W, b) := \frac{1}{n} \sum_{i=1}^n f_i(W, b) \right\}, \quad (9)$$

where  $f_i(W, b) = \ell(x_i, y_i, W, b) + \lambda P(W)$ . The function  $\eta_\alpha(t)$  can be expressed as a DC function:  $\eta_\alpha(t) = \alpha|t| - h(t)$ , where  $h(t) = -1 + \alpha|t| + \exp(-\alpha|t|)$ . Hence, the objective function  $f$  is a sum of  $n$  DC functions. We propose a special DC decomposition of  $f_i$  by moving  $\ell(x_i, y_i, W, b)$  into the second DC component as follows:  $f_i = g_i - h_i$ , where  $g_i$  and  $h_i$  are respectively given by

$$g_i(W, b) = \frac{\rho}{2} \|(W, b)\|^2 + \lambda \alpha \sum_{j=1}^d \|W_{:,j}\|_\infty,$$

$$h_i(W, b) = \frac{\rho}{2} \|(W, b)\|^2 - \ell(x_i, y_i, W, b) + \lambda \sum_{j=1}^d h(\|W_{j,:}\|_\infty),$$

are convex functions for some large enough  $\rho$ . This decomposition leads to a simple stochastic DCA scheme in which the convex optimization sub-problem can be efficiently solved. We note that  $\ell(x_i, y_i, W, b)$  is differentiable with  $L$ -Lipschitz gradient. Hence, we can choose  $\rho > L$ .

According to stochastic DCA scheme, at each iteration  $l$  we randomly choose one index  $i_l \in \{1, \dots, n\}$ , compute  $V^l \in \partial h_{i_l}(W^l, b^l)$  and then compute  $(W^{l+1}, b^{l+1})$  as the solution to the following convex problem

$$\min \left\{ \frac{1}{l+1} \sum_{h=0}^l [g_{i_h}(W, b) - \langle V^h, (W, b) \rangle] \right\}. \quad (10)$$

Let  $\text{softmax}(b^l + (W^l)^T x_i), z_i \in \mathbb{R}^Q$  be vectors respectively defined by

$$\begin{aligned} \text{softmax}(b^l + (W^l)^T x_i)_k &= \frac{\exp(b_k^l + (W_{:,k}^l)^T x_i)}{\sum_{h=1}^Q \exp(b_h^l + (W_{:,h}^l)^T x_i)}, \\ z_{ik} &= 1 \text{ if } k = y_i \text{ and } 0 \text{ otherwise,} \end{aligned}$$

for  $k = 1, \dots, Q$ . We have  $V^l = \rho(W^l, b^l) - (\nabla_W^l, \nabla_b^l) + (Y^l, 0)$ , where

$$\begin{cases} \nabla_W^l &= x_{i_l} (\text{softmax}(b^l + (W^l)^T x_{i_l}) - z_{i_l})^T, \\ \nabla_b^l &= \text{softmax}(b^l + (W^l)^T x_{i_l}) - z_{i_l}, \end{cases} \quad (11)$$

The computation of  $Y^l$  is described by

$$Y_{j,k}^l = \begin{cases} 0 & \text{if } k = k_0^j, \\ \lambda \alpha \eta_\alpha (\|W_{j,:}^l\|_\infty) \text{sign}(W_{j,k}^l) & \text{otherwise,} \end{cases} \quad (12)$$

where  $k_0^j = \arg \max_k |W_{j,k}^l|$ . The convex problem (10) can be separated into the following two problems

$$\min \left\{ \frac{1}{2} \|b\|_2^2 - \left\langle \frac{1}{l+1} \sum_{h=0}^l b^h + \nabla_b^h / \rho, b \right\rangle \right\}, \quad (13)$$

$$\min \left\{ \frac{1}{2} \|W\|^2 + \frac{\lambda \alpha}{\rho} \sum_{j=1}^d \|W_{j,:}\|_\infty - \langle R^l, W \rangle \right\}, \quad (14)$$

where  $R^l = \frac{1}{l+1} \sum_{h=0}^l [W^h - \nabla_W^h / \rho + Y^h / \rho]$ . The solution to the problem (13) is given by

$$b^{l+1} = b^l - \frac{1}{\rho(l+1)} \nabla_b^l. \quad (15)$$

We observe that the objective of the problem (14) is separable in rows of  $W$ , then the solution to this problem can be computed by solving  $d$  independent sub-problems

$$\min \left\{ \frac{1}{2} \|W_{j,:}\|^2 + \frac{\lambda \alpha}{\rho} \|W_{j,:}\|_\infty - \langle R_{j,:}^l, W_{j,:} \rangle \right\}, \quad (16)$$

where  $R_{j,:}^l = \frac{1}{l+1} \sum_{h=0}^l [W_{j,:}^h - \nabla_{W_{j,:}}^h / \rho + Y_{j,:}^h / \rho]$ . The solution to this problem is computed by the following operator proximal

$$W_{j,:}^{l+1} = \mathbf{prox}_{\lambda\alpha/\rho\|\cdot\|_\infty} (R_{j,:}^l). \quad (17)$$

By Moreau decomposition, it follows that

$$\mathbf{prox}_{\gamma\|\cdot\|_\infty} (v) = v - \gamma \mathbf{proj}_{\mathcal{B}} (v/\gamma), \quad (18)$$

where  $\mathcal{B} = \{x : \|x\|_1 \leq 1\}$  is the unit  $\ell_1$  ball and  $\mathbf{proj}_{\mathcal{B}}$  is the projection onto  $\mathcal{B}$ . We have

$$\mathbf{proj}_{\mathcal{B}} (v/\gamma) = \begin{cases} v/\gamma & \text{if } \|v\|_1 \leq \gamma, \\ \text{sign}(v) \cdot (|v/\gamma| - \delta)_+ & \text{if } \|v\|_1 > \gamma, \end{cases} \quad (19)$$

where  $\delta$  is computed as the solution to  $\sum_{k=1}^Q (|v_k/\gamma| - \delta)_+ = 1$ . Applying the above decomposition, the solution to the problem (16) is computed by

$$W_{j,:}^{l+1} = \begin{cases} 0 & \text{if } \|R_{j,:}^l\|_1 \leq \frac{\lambda\alpha}{\rho} \\ R_{j,:}^l - \text{sign}(R_{j,:}^l) \cdot (|R_{j,:}^l| - \delta)_+ & \text{if } \|R_{j,:}^l\|_1 > \frac{\lambda\alpha}{\rho}, \end{cases} \quad (20)$$

where  $\delta$  is computed as the solution to

$$\sum_{k=1}^Q (|R_{j,k}^l| - \delta)_+ = \frac{\lambda\alpha}{\rho}. \quad (21)$$

For computing  $\delta$  in (21), some efficient algorithms are available. Among them, we use the very inexpensive algorithm developed in [3]. Hence, the stochastic DCA for solving the problem (9) is described as follows.

---

**SDCA** (Stochastic DCA for solving the problem (9))

---

**Initialization:** Choose  $W^0 \in \mathbb{R}^{d \times Q}$ ,  $b^0 \in \mathbb{R}^Q$  and  $\rho > L$ .

**For**  $l = 0, 1, \dots$  **do**

1. Randomly choose an index  $i_l \in \{1, \dots, n\}$ .
2. Compute  $\nabla_{W}^l$ ,  $\nabla_b^l$  and  $Y^l$  using (11)-(12).
3. Compute  $(W^{l+1}, b^{l+1})$  via (15) and (20).

**End for.**

---

## 3 Numerical Experiment

### 3.1 Datasets

To study the performances of algorithms, we performed numerical experiments on 2 types of data: synthetic datasets (*sim\_1*, *sim\_2* and *sim\_3*) and real-world

datasets (*aloi*, *covertype* and *sensorless*). *aloi* [4] is a collection of 1,000 objects from various imaging circumstances, with a total of 110,250 images where each object is recorded over 100 images. *covertype* belongs to the Forest Cover Type Prediction from strictly cartographic variables. This is a very large dataset containing 581,012 points described by 54 features. *sensorless*<sup>1</sup> is a large dataset, which contains 58,509 data points, described by 48 features and classified into 11 classes.

For synthetic datasets, we generated *sim\_1*, *sim\_2* and *sim\_3* as proposed in [22]. In the first dataset (*sim\_1*), features are independent and have different averages in each class. In dataset *sim\_2*, features also have different averages in each class, but they are dependent. The last simulated dataset (*sim\_3*) has different one-dimensional averages in each class with independent features. The procedure for generating simulated datasets is described as follows:

- For *sim\_1*: we generate a four classes classification problem. Each class is assumed to have a multivariate normal distribution  $\mathcal{N}(\mu_k, I)$ ,  $k = 1, 2, 3, 4$  with dimension of  $d = 50$ . The first 10 components of  $\mu_1$  are 0.5,  $\mu_{2j} = 0.5$  if  $11 \leq j \leq 20$ ,  $\mu_{3j} = 0.5$  if  $21 \leq j \leq 30$ ,  $\mu_{4j} = 0.5$  if  $31 \leq j \leq 40$  and 0 otherwise. We generate 100,000 instances.
- For *sim\_2*: this simulation includes three classes of multivariate normal distributions  $\mathcal{N}(\mu_1, \Sigma)$ ,  $\mathcal{N}(\mu_2, \Sigma)$  and  $\mathcal{N}(\mu_3, \Sigma)$  each of dimension  $d = 50$ . The components of  $\mu_1$  are assumed to be 0,  $\mu_{2j} = 0.4$  and  $\mu_{3j} = 0.8$  if  $j \leq 40$  and 0 otherwise. The covariance matrix  $\Sigma$  is the block diagonal matrix with five blocks of dimension  $10 \times 10$  whose element  $(j, j')$  is  $0.6^{|j-j'|}$ . For each class, we generate 50,000 instances.
- For *sim\_3*: we generate a four-class classification problem as follows:  $i \in C_k$  then  $X_{ij} \sim \mathcal{N}((k-1)/3, 1)$  if  $j \leq 100$ ,  $k = 1, 2, 3, 4$  and  $X_{ij} \sim \mathcal{N}(0, 1)$  otherwise, where  $\mathcal{N}(\mu, \sigma^2)$  denotes the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . A total of 250,000 instances are generated with equal probabilities for each class.

The information about both real and synthetic datasets are summarized in the first column of Table 1.

### 3.2 Comparative Algorithms

We compare our algorithm with two others. The first one is **LibLinear**<sup>2</sup>, a well-known package in solving large-scale problems by using the coordinate descent algorithm. **LibLinear** contains several logistic regression classifiers, and we use **LibLinear**'s  $\ell_1$ -regularized logistic regression solver for this problem. In detail, it solves the following binary classification problem:

$$\min_w \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i}) + \lambda \sum_{j=1}^d |w_j|.$$

Then, one-vs-the-rest strategy is used for the multiclass problem.

<sup>1</sup> <https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis>.

<sup>2</sup> <https://cran.r-project.org/web/packages/LiblinearR/index.html>.

The second comparative algorithm is **GLASSO** [5], which is implemented as a R package<sup>3</sup>. **GLASSO** is a method for variable selection which solves the following  $\ell_1$  constraint generalized LASSO models:

$$\min_{W,b} \sum_{i=1}^n l(x_i, y_i, W, b) \quad \text{s.t.} \quad \sum_{i=1}^Q \sum_{j=1}^d |W_{j,i}| \leq \lambda.$$

### 3.3 Experiment Setting

To evaluate the performance of algorithms, we consider three criteria: classification accuracy, percentage of selected features (sparsity) and CPU time (measured in seconds). Sparsity is computed as the percentage of selected features, where a feature  $j \in \{1, \dots, d\}$  is considered to be removed if  $|W_{j,i}| < 10^{-8}, \forall i \in 1, \dots, Q$ .

For cross-validation process, the dataset is randomly split by the ratio of 80%–20% into a training set and a test set. Each algorithm will use the training set to learn its decision function. Classification accuracy is reported by evaluating on test set, whereas CPU time and sparsity are both reported based on training process. This process is repeated 10 times. We report the mean and standard deviation of each criteria.

Stopping condition of **SDCA** is determined by using a validation set, which is split randomly from the training set by the ratio of 80%–20%. After each epoch, the  $\text{accuracy}_{\text{valid}}$  is computed as accuracy on the validation set. If the  $\text{accuracy}_{\text{valid}}$  does not improve after  $n_{\text{patience}} = 5$  epochs then **SDCA** stops. For comparative algorithms, their stopping parameters are set as default. We also stop algorithms if their training process exceed 2h of CPU time.

For **SDCA**, we set the trade-off parameter  $\lambda \in \{10^{-4}, \dots, 1\}$ , while for **LibLinear** and **GLASSO** it is chosen in  $\{10^{-3}, \dots, 10^4\}$ . The parameter for controlling the tightness of zero-norm approximation  $\alpha$  is taken in  $\{0.5, 1, 2, 5\}$ .

For pre-processing data, we use standardization to scale the data. Features which have standard deviation lower than  $10^{-8}$  are removed.

All experiments are performed on a PC Intel (R) Xeon (R) E5-2630 v2 @2.60 GHz of 32 GB RAM on Windows 7.

### 3.4 Experimental Results on Synthetic Datasets

For synthetic datasets (*sim\_1*, *sim\_2* and *sim\_3*), we have known in advance about informative features that these datasets are generated from. The first experiment’s purpose is to examine algorithm’s feature selection ability, whether or not they can select informative features to provide high classification accuracy.

We obtain the comparative results as in Table 1 and we observe that:

- For all three synthetic datasets, **SDCA** successfully selects the exact informative features. **LibLinear** selects the exact features on 2 out of 3 datasets while **GLASSO** succeeds on only 1 dataset.

<sup>3</sup> <https://stat.snu.ac.kr/ydkim/programs/glasso/index.html>.



**Table 1.** Comparative results on both real and synthetic datasets. Bold values correspond to best results for each dataset. *NA* means that the algorithm fails to furnish a result.  $n$ ,  $d$  and  $Q$  is the number of instances, the number of dimensions and the number of classes respectively.

Dataset	Algorithm	Accuracy (%)		Time (sec)		Sparsity (%)	
		Mean	STD	Mean	STD	Mean	STD
sim_1 ( $n \times d$ ) = (100,000 $\times$ 50) Q = 4	SDCA	72.23	0.44	<b>1.08</b>	0.22	<b>80.00</b>	0.00
	LibLinear	<b>72.62</b>	0.38	2038.85	5.05	<b>80.00</b>	0.00
	GLASSO	72.49	0.15	2047.67	8.04	<b>80.00</b>	0.00
sim_2 ( $n \times d$ ) = (150,000 $\times$ 50) Q = 3	SDCA	<b>68.51</b>	0.16	<b>1.37</b>	0.45	80.00	0.00
	LibLinear	66.92	0.02	2.04	0.23	80.00	0.00
	GLASSO	62.51	0.58	2821.60	724.31	<b>74.67</b>	1.15
sim_3 ( $n \times d$ ) = (250,000 $\times$ 500) Q = 4	SDCA	<b>99.93</b>	0.01	<b>34.34</b>	5.00	<b>80.00</b>	0.00
	LibLinear	99.03	0.00	50.50	2.96	97.16	0.50
	GLASSO	NA	NA	NA	NA	NA	NA
aloi ( $n \times d$ ) = (108,000 $\times$ 128) Q = 1,000	SDCA	<b>85.61</b>	0.35	<b>112.77</b>	22.73	<b>100.00</b>	0.00
	LibLinear	81.61	0.20	2732.96	46.38	<b>100.00</b>	0.00
	GLASSO	NA	NA	NA	NA	NA	NA
coverttype ( $n \times d$ ) = (581,012 $\times$ 54) Q = 7	SDCA	<b>71.56</b>	0.25	<b>14.83</b>	4.19	<b>87.04</b>	2.62
	LibLinear	71.54	0.19	264.88	26.83	100.00	0.00
	GLASSO	NA	NA	NA	NA	NA	NA
sensorless ( $n \times d$ ) = (58,509 $\times$ 48) Q = 11	SDCA	<b>84.23</b>	0.99	<b>1.87</b>	0.09	100.00	0.00
	LibLinear	75.55	0.24	216.48	72.05	100.00	0.00
	GLASSO	53.59	0.73	2383.24	18.41	<b>91.67</b>	4.17

- Regarding to classification accuracy, SDCA gives the highest accuracy on both *sim\_2* and *sim\_3* dataset, and only 0.4% lower than the best algorithm (LibLinear) on *sim\_1* dataset.
- In terms of CPU time, SDCA is the fastest algorithm in this experiment. SDCA is up to 1887 times (resp. 2059 times) faster than LibLinear (resp. GLASSO). GLASSO fails in *sim\_3* dataset by overrun the 2 h limit.

### 3.5 Experimental Results on Real Datasets

In this experiment, we perform the comparative study on real datasets, namely *aloi*, *coverttype* and *sensorless*. The comparative results are reported in Table 1. We observe that:

- For *aloi* dataset, SDCA is the best algorithm among the three. Although both SDCA and LibLinear select all features, SDCA achieves the accuracy of 85.61%, which is 4% higher than LibLinear. Moreover, SDCA is 24.23 times faster than LibLinear while GLASSO fails to furnish a result after 2 h of CPU time.
- For *coverttype* dataset, SDCA gives a slightly better classification accuracy (71.56%) comparing to LibLinear (71.54%). Moreover, LibLinear fails to

suppress features while **SDCA** selects 87.04% of features. Concerning the CPU time, **SDCA** is 17.86 times faster than **LibLinear**. Hence **SDCA** is better than **LibLinear** in all three aspects. **GLASSO** fails again to provide a solution within the limitation of 2 h of CPU time.

- For *sensorless* dataset, **SDCA** is better than both **LibLinear** and **GLASSO** on both classification accuracy and running time. In term of accuracy, the gains versus **LibLinear** (resp. **GLASSO**) is high, which is 8.68% (resp. 30.64%). Regarding to running time, **SDCA** is 1274 times faster than **GLASSO** and 115 times faster than **LibLinear**. Although **GLASSO** is able to suppress 9.33% of all features, both classification accuracy and running time of **GLASSO** are worse than both **SDCA** and **LibLinear** by a big margin.

Overall, on all three real-world datasets, **SDCA** is the best among three comparative algorithms on both classification accuracy and CPU time. In terms of sparsity, **SDCA** outperforms the other two algorithms on 2 out of 3 datasets.

## 4 Conclusion

We have rigorously studied the DC programming and DCA for the sparse multiclass logistic regression problem. Using the  $\ell_{\infty,0}$  regularization, the resulting optimization problem is non-convex. The  $\ell_{\infty,0}$ -norm is approximated by a continuous function based on the piecewise exponential function. The latter is then reformulated as a DC function and we developed a stochastic version of DCA to solve it. This approach can handle datasets with very large number of instances. At each iteration, our stochastic algorithm only uses a small subset of data instances. By exploiting the structure of the problem, we propose an efficient DC decomposition for which the corresponding stochastic DCA scheme is very inexpensive.

Numerical experiments were carefully conducted on both synthetic and real-world datasets. The numerical results show that our algorithm **SDCA** outperforms well-known algorithms (**Liblinear** and **GLASSO**) in term of classification accuracy, sparsity of solution and running time. Especially, the gain on running time is huge. We are convinced that stochastic DCA is a promising approach for handling very large-scale datasets in machine learning.

## References

1. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theor. Comput. Sci.* **209**(1), 237–260 (1998)
2. Cox, D.: The regression analysis of binary sequences (with discussion). *J. Roy. Stat. Soc. B* **20**, 215–242 (1958)
3. Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T.: Efficient projections onto the  $l_1$ -ball for learning in high dimensions. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279. ACM (2008)
4. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.: The Amsterdam library of object images. *Int. J. Comput. Vis.* **61**(1), 103–112 (2005)

5. Kim, J., Kim, Y., Kim, Y.: A gradient-based optimization algorithm for LASSO. *J. Comput. Graph. Stat.* **17**(4), 994–1009 (2008)
6. King, G., Zeng, L.: Logistic regression in rare events data. *Polit. Anal.* **9**, 137–163 (2001)
7. Le, H.M., Le Thi, H.A., Nguyen, M.C.: Sparse semi-supervised support vector machines by DC programming and DCA. *Neurocomputing* **153**, 62–76 (2015)
8. Le Thi, H.A., Le, H.M., Nguyen, V.V., Pham Dinh, T.: A DC programming approach for feature selection in support vector machines learning. *Adv. Data Anal. Classif.* **2**(3), 259–278 (2008)
9. Le Thi, H.A., Nguyen, M.C.: Efficient algorithms for feature selection in multi-class support vector machine. In: *Advanced Computational Methods for Knowledge Engineering*, pp. 41–52. Springer, Heidelberg (2013)
10. Le Thi, H.A., Nguyen, T.B.T., Le, H.M.: Sparse signal recovery by difference of convex functions algorithms. In: *Intelligent Information and Database Systems*, pp. 387–397. Springer, Heidelberg (2013)
11. Le Thi, H.A., Pham Dinh, T.: The DC (Difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization Problems. *Ann. Oper. Res.* **133**(1–4), 23–46 (2005)
12. Le Thi, H.A., Pham Dinh, T., Le, H.M., Vo, X.T.: DC approximation approaches for sparse optimization. *Eur. J. Oper. Res.* **244**(1), 26–46 (2015)
13. Le Thi, H.A., Phan, D.N.: DC Programming and DCA for Sparse Optimal Scoring Problem. *Neurocomput.* **186**(C), 170–181 (2016)
14. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Math. Vietnamica* **22**(1), 289–355 (1997)
15. Pham Dinh, T., Le Thi, H.A.: A D.C. Optimization algorithm for solving the trust-region subproblem. *SIAM J. Optim.* **8**(2), 476–505 (1998)
16. Phan, D.N.: Algorithmes basés sur la programmation DC et DCA pour l'apprentissage avec la parcimonie et l'apprentissage stochastique en grande dimension. Université de Lorraine, Thèse de doctorat (2016)
17. Quattoni, A., Carreras, X., Collins, M., Darrell, T.: An Efficient Projection for L1,  $\infty$  Regularization. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 857–864. ICML 2009. ACM, New York (2009)
18. Turlach, B.A., Venables, W.N., Wright, S.J.: Simultaneous variable selection. *Technometrics* **47**(3), 349–363 (2005)
19. Verma, J.P.: Logistic regression: developing a model for risk analysis. In: *Data Analysis in Management with SPSS Software*, pp. 413–442. Springer, India (2013)
20. Wang, L., Chen, G., Li, H.: Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**(12), 1486–1494 (2007)
21. Wei, F., Zhu, H.: Group coordinate descent algorithms for nonconvex penalized regression. *Comput. Stati. Data Anal.* **56**(2), 316–326 (2012)
22. Witten, D.M., Tibshirani, R.: Penalized classification using fisher's linear discriminant. *J. Roy. Stat. Soc. B* **73**(5), 753–772 (2011)
23. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B* **68**, 49–67 (2006)
24. Zhang, H.H., Liu, Y., Wu, Y., Zhu, J.: Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electron. J. Stat.* **2**, 149–167 (2008)

# Reformulation of the Quadratic Multidimensional Knapsack Problem as Copositive/Completely Positive Programs

D.V. Nguyen<sup>(✉)</sup>

Department of Mathematics, University of Trier, 54286 Trier, Germany  
duy\_van@gmx.de

**Abstract.** The general (nonconvex) quadratic multidimensional knapsack problem (QMKP) is one of the most important combinatorial optimization problems with many practical applications. The purpose of this article is to establish equivalent formulations of (QMKP) as so called copositive programs and completely positive programs. The resulting programs can then be handled by copositive programming methods, which are completely different from classical algorithms for directly solving quadratic knapsack problems.

**Keywords:** Knapsack problem · Quadratic multidimensional knapsack problem · Copositive programming · Completely positive program

## 1 Introduction

Let  $\mathbb{R}^d$  and  $\mathbb{R}_+^d$  be the  $d$  dimensional real space and its nonnegative orthant, respectively,  $\mathbb{R}^{d \times k}$  the space of  $d \times k$  matrices, and

$$\mathcal{S}_d := \{S \in \mathbb{R}^{d \times d} \mid S^T = S\}$$

the space of symmetric matrices.

The subject of this article is the general (nonconvex) quadratic multidimensional knapsack problem, which is formally formulated as follows (see [7, 16]).

$$(QMKP) \begin{cases} \min x^T \bar{Q} x + \bar{c}^T x \\ \text{s.t. } \bar{a}_i^T x \leq \bar{b}_i & \text{for all } i = 1, \dots, m \\ x \in \{0, 1\}^d, \end{cases}$$

where  $\bar{Q} \in \mathcal{S}_d$ ,  $\bar{c} \in \mathbb{R}^d$ ,  $\bar{a}_i \in \mathbb{R}_+^d$  and  $\bar{b}_i > 0$  for all  $i = 1, \dots, m$ .

A well known special case of (QMKP) is the quadratic knapsack problem, (QKP), containing only one capacity constraint, i.e., the case where  $m = 1$ .

One of the most important applications of Problem (QKP) is the portfolio management problem, which can be formulated as an optimization problem with a quadratic objective function under a knapsack constraint (see, e.g., [11]).

The quadratic function measures both the expected return and the risk. The single knapsack constraint represents the budget restriction. By using only one knapsack constraint, it is not allowed to consider the possibility of investing into assets of different risk levels. Therefore, several knapsack constraints should be considered, each of them represents a budget allocated to assets of a given risk level. More details about this capital budgeting model can be found in Faaland [9] and Djerdjour et al. [7].

While the Quadratic Knapsack Problem (QKP) is a much-studied combinatorial optimization problem, (see, e.g. [2,3,13] and a survey of Pisinger in [14] and references given therein), there are only a few methods for handling some specific cases of the quadratic multidimensional knapsack problem (QMKP), see e.g., [7,15,16].

In the last two decades, a relatively young field in mathematical optimization called copositive programming has received a great deal of attention from researchers. This is a class of linear programs with matrix variables and additional conic constraints defined by the cones of copositive or completely positive matrices. It has been shown that there is a close relationship between (continuous or binary) quadratic problems and copositive programs, see e.g. [4–6,8,10,12] and references given therein.

The purpose of this paper is to establish equivalent formulations of (QMKP) as copositive programs and completely positive programs. The resulting problems can then be handled by solution methods (see, e.g., [5,8,12] and references given therein), which are completely different from known algorithms for directly solving quadratic knapsack problems.

In the next section we give some preliminaries on copositive programs and completely positive programs. Equivalent Reformulations of (QMKP) as copositive programs and completely positive programs are established thereafter.

## 2 Preliminaries

### 2.1 Copositive and Completely Positive Cones

**Definition 1.** Let  $A$  be a  $d \times d$  real symmetric matrix. One says that  $A$  is copositive if  $x^T A x \geq 0$  for all  $x \geq 0$ . Strict copositivity of  $A$  means that  $x^T A x > 0$  for all  $x \geq 0$ ,  $x \neq 0$ .

Let  $\mathcal{COP}_d$  be the set of all  $d \times d$  copositive matrices. Then (see, e.g., [1,8,10])  $\mathcal{COP}_d$  is a closed convex pointed cone in the space of symmetric matrices  $\mathcal{S}_d$  with  $\text{int}(\mathcal{COP}_d) \neq \emptyset$ .

**Definition 2.** Let  $A$  be a  $d \times d$  real symmetric matrix. One says that  $A$  is completely positive if there exists an integer  $m$  and a  $d \times m$  matrix  $B$  with non-negative entries such that  $A = B B^T$ . The smallest possible number  $m$  is called the CP-rank of  $A$ .

Let  $\mathcal{CP}_d$  be the set of all  $d \times d$  completely positive matrices. Then  $\mathcal{CP}_d$  is a closed convex pointed cone in  $\mathcal{S}_d$  with  $\text{int}(\mathcal{CP}_d) \neq \emptyset$  (see, e.g., [1,8,10]).

**Definition 3.** Let  $\mathcal{C}$  be an arbitrary given cone in  $\mathcal{S}_d$ , the dual cone  $\mathcal{C}^*$  to  $\mathcal{C}$  is defined as

$$\mathcal{C}^* = \{A \in \mathcal{S}_d \mid \langle A, B \rangle \geq 0 \forall B \in \mathcal{C}\},$$

where

$$\langle A, B \rangle = \text{tr}(A^T B) = \sum_{i=1}^d \sum_{j=1}^d a_{ij} b_{ij};$$

It is well known (see, e.g., [1, 8, 10], and references given therein) that the cones  $\mathcal{COP}_d$  and  $\mathcal{CP}_d$  are dual to each other in the sense that

$$\mathcal{COP}_d^* = \mathcal{CP}_d \text{ and } \mathcal{CP}_d^* = \mathcal{COP}_d. \quad (1)$$

## 2.2 Copositive and Completely Positive Programs and Their Duals

Let  $Q \in \mathcal{S}_d$ ,  $A_i \in \mathcal{S}_d$ ,  $b_i \in \mathbb{R}$ ,  $i = 1, \dots, m$ , and  $\mathcal{K}$  a convex cone in  $\mathcal{S}_d$ . Consider a linear optimization problem in matrix variables with a conic constraint of the following form:

$$\begin{aligned} \min \quad & \langle Q, X \rangle \\ \text{s.t.} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \dots, m \\ & X \in \mathcal{K}. \end{aligned} \quad (2)$$

**Definition 4.** Problem (2) is called *copositive program* if  $\mathcal{K} = \mathcal{COP}_d$ . It is called *completely positive program* if  $\mathcal{K} = \mathcal{CP}_d$ .

The corresponding Lagrangian dual of Problem (2) is then

$$\begin{aligned} \max \quad & \sum_{i=1}^m b_i y_i \\ \text{s.t.} \quad & Q - \sum_{i=1}^m y_i A_i \in \mathcal{K}^* \\ & y_i \in \mathbb{R}, \quad i = 1, \dots, m. \end{aligned} \quad (3)$$

Since  $\mathcal{K}$  and  $\mathcal{K}^*$  are convex cones, the strong duality requires some constraint qualifications such as Problem (2) respectively Problem (3) to be strict feasible, i.e., there exists a feasible point in  $\text{int}(\mathcal{K})$  or  $\text{int}(\mathcal{K}^*)$ , respectively.

## 2.3 Quadratic Optimization Problems and Completely Positive Programs

There exists a close relationship between quadratic optimization problems and completely positive/copositive programs. We discuss this relationship by the following two known cases.

First, consider the so-called standard quadratic optimization problem in [4]:

$$\begin{aligned} \min \quad & x^T Q x \\ \text{s.t.} \quad & e^T x = 1 \\ & x \geq 0, \end{aligned} \quad (4)$$

where  $Q \in \mathcal{S}_d$  and  $e$  denotes the all-ones vector. The authors of [4] showed that Problem 4 is equivalent to the following completely positive program:

$$\begin{aligned} \min \quad & \langle Q, X \rangle \\ \text{s.t.} \quad & \langle ee^T, X \rangle = 1 \\ & X \in \mathcal{CP}_d. \end{aligned} \quad (5)$$

The equivalence between Problem 4 and Problem 5 is that each extreme optimal solution of 5 has the form  $X^* = x^*(x^*)^T$ , where  $x^*$  is an optimal solution of Problem 4 and both problems have the same optimal values.

The second problem is the mixed-binary quadratic program considered by Burer in [6],

$$\begin{aligned} \min \quad & x^T Q x + 2q^T x \\ \text{s.t.} \quad & a_i^T x = b_i, \quad i = 1, \dots, m \\ & x \geq 0 \\ & x_j \in \{0, 1\}, \quad j \in J \subseteq \{1, \dots, d\}, \end{aligned} \quad (6)$$

where  $Q \in \mathcal{S}_d$ ,  $q, a_i \in \mathbb{R}^d$ ,  $i = 1, \dots, m$ .

Under the following two *Key Assumptions*:

(KA1) It holds:  $x \in L \implies 0 \leq x_j \leq 1, j = 1, \dots, d$ , where

$$L = \{x \geq 0 : a_i^T x = b_i, \quad i = 1, \dots, m\},$$

(KA2)  $\exists \beta \in \mathbb{R}^m$  such that

$$\sum_{i=1}^m \beta_i a_i^T \geq 0, \quad \sum_{i=1}^m \beta_i b_i = 1,$$

Burer [6] showed that under (KA1)–(KA2), by using a vector

$$\alpha = \sum_{i=1}^m \beta_i a_i \geq 0, \quad (7)$$

Problem (6) can then be equivalently reformulated as the following completely positive program:

$$\begin{aligned} \max \quad & \langle Q, X \rangle + 2q^T X \alpha \\ \text{s.t.} \quad & a_i^T X \alpha = b_i, \quad i = 1, \dots, m \\ & a_i^T X a_i = b_i^2, \quad i = 1, \dots, m \\ & (X \alpha)_j = X_{jj}, \quad j \in J \\ & \alpha^T X \alpha = 1 \\ & X \in \mathcal{CP}_d. \end{aligned} \quad (8)$$

The equivalence between Problem (6) and Problem (8) is stated as follows (see [6]):

**Proposition 1.** *Under (KA1)–(KA2), let  $\alpha$  be defined as in (7). Then Problem (6) is equivalent to Problem (8) in the sense that:*

- (i) *The optimal values of both problems are equal;*
- (ii) *If  $X^*$  is an optimal solution of Problem (8), then  $X^* \alpha$  lies in the convex hull of optimal solutions for Problem (6).*

### 3 Construction of Equivalent Completely Positive and Copositive Optimization Problems

Based on the reformulation principle developed in [12], we present two copositive/completely positive optimization problem models for Problem  $(QMKP)$ . Moreover, we also show that the resulting primal-dual pairs have strong duality property.

#### Model I

The first completely positive optimization problem model is constructed by the following steps.

*Step 1.* Add a redundant constraint  $\bar{a}^T x \leq 1$  with  $\bar{a} > 0$  to the system of linear inequalities. There are two simple ways to construct such a vector  $\bar{a}$ .

- (i) If there exists  $y \in \mathbb{R}_+^m$  with  $\sum_{i=1}^m y_i \bar{a}_i > 0$ . Then set

$$\bar{a} := \frac{1}{\bar{b}^T y} \sum_{i=1}^m y_i \bar{a}_i > 0,$$

where  $\bar{b}^T = (\bar{b}_1, \dots, \bar{b}_m) > 0$ .

- (ii) Set

$$\bar{a} := \frac{1}{n-1} e,$$

where  $e$  is the vector with all components equal to 1.

As a result, we have a slightly modified problem

$$(QMKP) \begin{cases} \min x^T \bar{Q} x + \bar{c}^T x \\ \text{s.t. } \bar{a}_i^T x \leq \bar{b}_i & \text{for all } i = 1, \dots, m \\ \bar{a}^T x \leq 1 \\ x \in \{0, 1\}^d. \end{cases}$$

*Step 2.* Add a slack variable  $s$  to the constraint  $\bar{a}^T x \leq 1$ , obtaining

$$(QMKP) \begin{cases} \min x^T \bar{Q} x + \bar{c}^T x \\ \text{s.t. } \bar{a}_i^T x \leq \bar{b}_i & \text{for all } i = 1, \dots, m \\ \bar{a}^T x + s = 1 \\ x \in \{0, 1\}^d \\ s \geq 0. \end{cases}$$

Define  $n := d + 1$  and

$$Q' := \begin{pmatrix} \bar{Q} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad c := \begin{pmatrix} \bar{c} \\ 0 \end{pmatrix}, \quad b := \begin{pmatrix} \bar{b}_1 \\ \vdots \\ \bar{b}_m \end{pmatrix} > 0,$$

$$a := \begin{pmatrix} \bar{a} \\ 1 \end{pmatrix} > 0, \quad A' := \begin{pmatrix} (\bar{a}_1^T, 0) \\ \vdots \\ (\bar{a}_m^T, 0) \end{pmatrix} \in \mathbb{R}_+^{m \times n}.$$



Then we obtain the following equivalent problem in  $\mathbb{R}^n$ , i.e., the vector of variables is now  $x \in \mathbb{R}^n$ .

$$(QMKP) \begin{cases} \min x^T Q' x + c^T x \\ \text{s.t. } A' x \leq b \\ a^T x = 1 \\ x_i \in \{0, 1\}, i = 1, \dots, n-1 \\ x_n \geq 0. \end{cases}$$

*Step 3.* (Getting rid of the linear term  $c^T x$  in the objective function) From the constraint  $a^T x = 1$  we have

$$c^T x = x^T a c^T x = \frac{1}{2} x^T (a c^T + c a^T) x.$$

Therefore, defining

$$Q := Q' + \frac{1}{2}(a c^T + c a^T),$$

we obtain

$$x^T Q' x + c^T x = x^T Q x.$$

Note that the matrix  $Q \in \mathbb{R}^{n \times n}$  is again symmetric.

Thus, using the trivial constraints

$$0 \leq x_j \leq 1 \text{ for all } j = 1, \dots, n-1,$$

we get the following equivalent problem.

$$(QMKP) \begin{cases} \min x^T Q x \\ \text{s.t. } A' x \leq b \\ 0 \leq x_j \leq 1 \text{ for all } j = 1, \dots, n-1 \\ a^T x = 1 \\ x_i \in \{0, 1\}, i = 1, \dots, n-1 \\ x_n \geq 0. \end{cases}$$

*Step 4.* Define matrix

$$A := (b)a^T - A'ea^T - I_{n-1}I_n \in \mathbb{R}^{(m+n-1+n) \times n}.$$

Then the above problem can be written as

$$(QMKP) \begin{cases} \min x^T Q x \\ a^T x = 1 \\ Ax \geq 0 \\ x_i \in \{0, 1\}, i = 1, \dots, n-1. \end{cases} \quad (9)$$

Note that for the last formulation of  $(QMKP)$  in (9) the following key assumption is fulfilled:

$$\text{For all } i = 1, \dots, n-1, \text{ we have } a_i > 0, \text{ and } 0 \leq x_i \leq 1 \text{ for all } x \in \mathbb{R}^n \text{ satisfying } Ax \geq 0, a^T x = 1. \quad (10)$$

*Step 5.* From the reformulation principle (Theorem 3.3 in [12]), the equivalent completely positive problem is

$$(CPMKP) \begin{cases} \min \langle Q, X \rangle \\ \langle aa^T, X \rangle = 1 \\ AXA^T \in \mathcal{CP}_{m+2n-1} \\ \langle B, X \rangle = 0, \end{cases}$$

where  $B := \sum_{i=1}^{n-1} e^i(a - e^i)^T$  with  $e^1, \dots, e^{n-1}$  being the first  $n - 1$  unit vectors of  $\mathbb{R}^n$ .

*Step 6.* The copositive program, which is the dual of  $(CPMKP)$  is

$$(COPMKP) \begin{cases} \max \lambda \\ Q - \lambda aa^T + sB + A^TUA \in \mathcal{COP}_n \\ A^TUA \in \mathcal{COP}_n \\ \lambda, s \in \mathbb{R} \\ U \in \mathcal{COP}_{m+2n-1}. \end{cases}$$

The strict feasibility of Problem  $(COPMKP)$  is shown as follows.

Choose  $\hat{U} = ee^T$ , so  $\hat{U} \in \text{int}\mathcal{COP}_{m+n}$ . Furthermore, choosing  $\hat{s} = 0$  and  $\hat{\lambda} < 0$  small enough, we obtain the matrix  $Q - \hat{\lambda}aa^T + A^T\hat{U}A$  with all positive entries, i.e.,

$$Q - \hat{\lambda}aa^T + A^T\hat{U}A \in \text{int } \mathcal{COP}_n.$$

## Model II

Consider again the quadratic multidimensional knapsack problem

$$(QMKP) \begin{cases} \min x^T \bar{Q}x + \bar{c}^T x \\ \text{s.t. } \bar{a}_i^T x \leq \bar{b}_i \quad \text{for all } i = 1, \dots, m, \\ x \in \{0, 1\}^{n-1}, \end{cases}$$

*Step 1.* Add constraints  $0 \leq x_j \leq 1$  for all  $j = 1, \dots, n - 1$  to the problem, obtaining

$$(QMKP) \begin{cases} \min x^T \bar{Q}x + \bar{c}^T x \\ \text{s.t. } \bar{a}_i^T x \leq \bar{b}_i \quad \forall i = 1, \dots, m, \\ x_j \leq 1 \quad \forall j = 1, \dots, n - 1 \\ x \in \{0, 1\}^{n-1} \\ x \geq 0. \end{cases}$$

*Step 2.* Add  $m + n - 1$  slack variables to transform the inequality constraints into equality constraints, obtaining

$$(QMKP) \begin{cases} \min x^T \bar{Q}x + \bar{c}^T x \\ \text{s.t. } \bar{a}_i^T x + s_i = \bar{b}_i \quad \forall i = 1, \dots, m, \\ x_j + s_{m+j} = 1 \quad \forall j = 1, \dots, n - 1 \\ x_j \in \{0, 1\} \quad \forall j = 1, \dots, n - 1 \\ x, s \geq 0. \end{cases}$$

*Step 3.* For each  $i = 1, \dots, m$ , set

$$a_i := \frac{1}{\bar{b}_i} \begin{pmatrix} \bar{a}_i \\ e_{m+n-1}^i \end{pmatrix},$$

where  $e_{m+n-1}^i \in \mathbb{R}^{m+n-1}$  is the  $i$ -th unit vector in  $\mathbb{R}^{m+n-1}$ .

For each  $j = 1, \dots, n-1$ , set

$$a_{m+j} := \begin{pmatrix} e_{n-1}^j \\ e_{m+n-1}^{m+j} \end{pmatrix},$$

where  $e_{n-1}^j \in \mathbb{R}^{n-1}$  is the  $j$ -th unit vector of  $\mathbb{R}^{n-1}$  and  $e_{m+n-1}^i \in \mathbb{R}^{m+n-1}$  is the  $i$ -th unit vector in  $\mathbb{R}^{m+n-1}$ .

Define

$$Q := \begin{pmatrix} \bar{Q} & O \\ O & O \end{pmatrix} \in \mathbb{R}^{(m+2n-2) \times (m+2n-2)} \text{ and } c := \begin{pmatrix} \bar{c} \\ 0 \end{pmatrix} \in \mathbb{R}^{m+2n-2}.$$

Then we can rewrite (QMKP) as a problem in  $\mathbb{R}^{m+2n-2}$  with the vector of variables  $y = \begin{pmatrix} x \\ s \end{pmatrix} \in \mathbb{R}^{(n-1)+(m+n-1)}$ , where  $s$  is the vector of slack variables:

$$(QMKP) \begin{cases} \min y^T Q y + c^T y \\ \text{s.t. } a_i^T y = 1 & \forall i = 1, \dots, m+n-1, \\ y_j \in \{0, 1\} & \forall j = 1, \dots, n-1 \\ y \geq 0. \end{cases}$$

*Step 4.* Adding a redundant equality constraint  $a^T y = 1$  with

$$a := \frac{1}{m+n-1} \sum_{i=1}^{m+n-1} \bar{a}_i,$$

we obtain

$$(QMKP) \begin{cases} \min y^T Q y + c^T y \\ \text{s.t. } a_i^T y = 1 & \forall i = 1, \dots, m+n-1, \\ y_j \in \{0, 1\} & \forall j = 1, \dots, n-1 \\ a^T y = 1 \\ y \geq 0. \end{cases}$$

Note that by construction, we have  $a > 0$ .

*Step 5.* Define following matrices:

$$\Omega := Q + \frac{1}{2}(ca^T + ac^T) \in \mathbb{R}^{(m+2n-2) \times (m+2n-2)},$$

$$C := (a)^T - a_1 : a^T - a_{m+n-1} \in \mathbb{R}^{(m+n-1) \times (m+2n-2)}$$

and

$$B := \sum_{i=1}^{n-1} e^i (a - e^i)^T + \sum_{i=1}^{n-1} (a - e^i) (e^i)^T \in \mathbb{R}^{(m+2n-2) \times (m+2n-2)},$$

where  $e^i$  is the  $i$ -th unit vector in  $\mathbb{R}^{m+2n-2}$ . Then Problem  $(QMKP)$  is lifted into the following completely positive problem:

$$(CPMKP) \begin{cases} \min \langle \Omega, Y \rangle \\ \text{s.t. } \langle CC^T, Y \rangle = 0 \\ \quad \langle B, Y \rangle = 0 \\ \quad \langle aa^T, Y \rangle = 1 \\ \quad Y \in \mathcal{CP}_{m+2n-2} \end{cases}$$

It is worth noting that, independently from the numbers of constraints and variables in Problem  $(QMKP)$ , Problem  $(CPMKP)$  has only two linear equality constraints and one completely positive constraint.

*Step 6.* The dual problem of  $(CPMKP)$  is the following copositive program:

$$(COPMKP) \begin{cases} \max \lambda \\ \text{s.t. } \Omega - \lambda aa^T + sCC^T + tB \in \mathcal{COP}_{m+2n-2} \\ \quad \lambda, s, t \in \mathbb{R} \end{cases}$$

A strictly feasible point of Problem  $(COPMKP)$  is easily determined as follows.

As the vector  $a$  defined in Step 4 satisfies  $a > 0$ , one can choose  $\hat{s} = 0$  and  $\hat{\lambda} < 0$  small enough, such that the matrix  $(\Omega - \hat{\lambda}aa^T)$  has all (strictly) positive entries, i.e. this matrix is an interior point of the copositive cone.

*Remark 1.* Using the formulation of Burer in [6] for the last problem  $(QMKP)$  constructed in Step 4, we obtain the following completely positive problem:

$$(CPMKP)_B \begin{cases} \min \langle Q, Y \rangle + c^T Y a \\ \text{s.t. } a_i^T Y a_i = 1 & \forall i = 1, \dots, m+n-1, \\ \quad a_i^T Y a = 1 & \forall i = 1, \dots, m+n-1 \\ \quad a^T Y a = 1 \\ \quad [Y a]_i = Y_{ii} & \forall i = 1, \dots, n-1 \\ \quad Y \in \mathcal{CP}_{m+2n-2}. \end{cases}$$

Note that this problem has  $2(m+n-1) + n$  linear equality constraints and one completely positive constraint. Thus, its dual problem has more variables than Problem  $(COPMKP)$  has. Moreover, it is also not clear, whether or not its dual problem is strictly feasible.

**Acknowledgment.** The author would like to thank two anonymous referees for their suggestions that help to improve the first version of this article.

## References

1. Berman, A., Shaked-Monderer, N.: *Completely Positive Matrices*. World Scientific, Singapore (2003)
2. Billionnet, A., Soutif, E.: Using a mixed integer programming tool for solving the 0–1 quadratic knapsack problem. *INFORMS J. Comput.* **16**, 188–197 (2004)
3. Billionnet, A., Soutif, E.: An exact method based on Lagrangian decomposition for the 0–1 quadratic knapsack problem. *Eur. J. Oper. Res.* **157**, 565–575 (2004)
4. Bomze, I.M., Dür, M., de Klerk, E., Roos, C., Quist, A.J., Terlaky, T.: On copositive programming and standard quadratic optimization problems. *J. Global Optim.* **18**, 301–320 (2000)
5. Bundfuss, S., Dür, M.: An adaptive linear approximation algorithm for copositive programs. *SIAM J. Optim.* **20**, 30–53 (2009)
6. Burer, S.: On the copositive representation of binary and continuous nonconvex quadratic programs. *Math. Program.* **120**, 479–495 (2009)
7. Djerdjour, M., Mathur, K., Salkin, H.: A surrogate-based algorithm for the general quadratic multidimensional knapsack. *Oper. Res. Lett.* **7**, 253–257 (1988)
8. Dür, M.: Copositive programming a survey. In: Diehl, M., Glineur, F., Jarlebring, E., Michiels, W. (eds.) *Recent Advances in Optimization and its Applications in Engineering*, pp. 3–20. Springer, Heidelberg (2010)
9. Faaland, B.: An integer programming algorithm for portfolio selection. *Manag. Sci.* **20**, 1376–1384 (1974)
10. Hiriart-Urruty, J.-B., Seeger, A.: A variational approach to copositive matrices. *SIAM Rev.* **52**, 593–629 (2010)
11. Markowitz, H.M.: Portfolio selection. *J. Finan.* **7**(1), 77–91 (1952)
12. Nguyen, D.V.: Contributions to quadratic optimization: algorithms, copositive programming reformulations and duality. Ph.D. thesis, Department of Mathematics, University of Trier (2017)
13. Pardalos, P.M., Ye, Y., Han, C.G.: Algorithms for the solution of quadratic knapsack problems. *Linear Algebra Appl.* **152**, 69–91 (1991)
14. Pisinger, D.: The quadratic knapsack problem—a survey. *Discret. Appl. Math.* **155**, 623–648 (2007)
15. Quadri, D., Soutif, E., Tolla, P.: Upper bounds for large scale integer quadratic multidimensional knapsack. *Int. J. Oper. Res.* **4**(3), 146–154 (2007)
16. Quadri, D., Soutif, E., Tolla, P.: Exact solution method to solve large scale integer quadratic multidimensional knapsack problems. *J. Comb. Optim.* **17**(2), 157–167 (2009)

# DC Programming and DCA for Enhancing Physical Layer Security in Amplify-and-Forward Relay Beamforming Networks Based on the SNR Approach

Nguyen Nhu Tuan<sup>(✉)</sup> and Dang Vu Son

Academy of Cryptography Techniques, Hanoi, Vietnam  
{nguyennhutuan,dvson}@bcy.gov.vn

**Abstract.** Physical layer security is an alternative approach to guarantee secure communication besides the traditional cryptography method. In the physical layer security literature, most of the studies are centered around the secrecy capacity, but a new approach based on signal-to-noise ratios (SNR) was proposed in recent years. In that approach, they considered the transmission of a confidential message over a wireless channel with the help of multiple cooperating relays in the Amplify-and-Forward cooperative scheme based on an SNR approach with the predefined threshold. The optimization problem with the aim of maximizing the received SNR at a legitimate receiver subjects to the conditions that the received SNR at each eavesdropper is below the target threshold, and the condition about relay power are formulated as a nonconvex problem. This problem was solved by Semi-definite Relaxation method. In this paper, we propose the well known method based on DC programming and DCA to solve this hard problem. The numerical results show that our proposed method give the better results compared with its obtained by the existing ones.

**Keywords:** DC programming and DCA · Amplify-and-Forward · Physical layer security

## 1 Introduction

In wireless networks, generally security measures are implemented in the higher layer of protocol stack using cryptographic methods. i.e., IP layer security (IPSec) with some common cryptography algorithms as 3DES, AES, GOST, etc. However, current advances in computation technology pose threats for such systems, prompting researchers to explore alternatives like physical layer security.

Physical layer security approach was pioneered by Wyner [14] in 1975 and subsequent works [1,6] have created interest in the information-theoretic aspects of physical layer security. In general, there are three main cooperative schemes in physical layer security, they are decode-and-forward (DF), amplify-and-forward (AF), and cooperative jamming (CJ). Most studies in all these

schemes focus on the secrecy capacity. The secrecy capacity is the maximum rate of information transfer from the transmitter to the receiver in the presence of eavesdropper(s). But, in [12], the authors have considered an SNR based model for secure communication. The motivation of their approach arises from the results that the secrecy capacity depends on the source-destination and the source-eavesdropper channel capacities [6], which in turn are function of the respective signal-to-noise ratios (SNR). The main principle of SNR approach is that the correct decoding of the received signals is dependent on the received SNR. If the received SNR of eavesdropper is below a certain predefined threshold, then for all practical purposes, we can say that the eavesdropper cannot extract any information from the received signal [12]. In this paper, we concentrate on the AF scheme, the other schemes are studied in our future works.

The remaining parts of this paper are arranged as follows. In the rest of this section (Sect. 1), we first shortly present the state-of-the-art of physical layer security in AF relay beamforming networks, and then briefly introduce DC programming and DCA. In Sect. 2, we describe the system model and formulated received SNR maximization problem. The existing work is mentioned in Sect. 3. Section 4 presents how to apply DC programming and DCA for solving the considered problem. Finally, Sects. 5 and 6 report numerical results and conclusions, respectively.

*Notation:* Let  $(\cdot)^T$ ,  $(\cdot)^\dagger$  and  $(\cdot)^*$  denote transpose, conjugate transpose and conjugate, respectively;  $\mathbf{I}_m$  is the identity matrix of size  $m \times m$ ;  $E\{\cdot\}$  denotes expectation;  $\langle \cdot, \cdot \rangle$  denotes the inner product and  $\|\cdot\|$  denotes the Euclidean norm. Bold lowercase letters denote column vectors and bold uppercase letters denote matrices.  $\mathbf{D}(\mathbf{a})$  denotes the diagonal matrix with  $\mathbf{a}$  on its main diagonal.  $\text{Re}(\cdot)$  and  $\text{Im}(\cdot)$  extract the real part and the imaginary part of its argument, respectively.

## 1.1 Physical Layer Security in AF Beamforming Protocol

AF cooperative scheme for improving the quality and transmission rate in the absence of any wire-tapper was studied in many works several years ago, but this cooperative scheme for improving communication secrecy rate in the presence of an or more eavesdroppers were considered in recent years ([7, 12, 13, 15]). For AF relay protocol, the secrecy capacity for a single eavesdropper, and the zero-forcing secrecy rate for multiple eavesdroppers under a total relay power constraint are evaluated in [2]. The techniques for computing the maximum AF secrecy rate with single and multiple eavesdroppers under both individual and total relay power constraints are devised in [7, 15]. However, in [12], the authors considered an SNR based model for secure communication. It arises from that the secrecy capacity depends on the channel capacities from source to destination, and from the source to eavesdropper, which in turn are functions of the respective signal-to-noise ratios.

In practical receiver, the correct decoding of the received is dependent on the received SNR. If the received SNR is below a certain predefined threshold, then

for all practical purposes, we can assume that the receiver cannot extract any information from the received signal. Following this argument, the authors in [12] considered the model where every eavesdropper has limited decoding capability, determined by its SNR threshold. Therefore, to ensure secure communication between the source and the destination, we strive to bring down the received SNR at each eavesdropper is below its target threshold.

The maximizing received SNR at the destination under received SNR at each eavesdropper constraints problem is also nonconvex and hard to solve. The parts below show the existing method ([11, 12]) and the proposed method based on DC programming and DCA for solving that problem.

## 1.2 DC Programming and DCA

DC (Difference of Convex functions) programming and DCA (DC Algorithm) were introduced by T. Pham Dinh in 1985 in their preliminary form and have been extensively developed by H.A. Le Thi and T. Pham Dinh since 1994 to become now classic and more and more popular. DCA is a continuous primal dual subgradient approach. It is relied on local optimality and duality in DC programming to solve standard DC programs which are of the form

$$\alpha = \inf\{f(x) := g(x) - h(x) : x \in \mathbb{R}^n\}, \quad (P_{dc})$$

with  $g, h \in \Gamma_0(\mathbb{R}^n)$ , which is a set of lower semi-continuous proper convex functions on  $\mathbb{R}^n$ . Such a function  $f(x)$  is called a DC function, and  $g(x) - h(x)$  is a DC decomposition of  $f(x)$ , while the convex functions  $g(x)$  and  $h(x)$  are DC components of  $f(x)$ . Any constrained DC program whose feasible set  $C$  is convex could be transformed into an unconstrained DC program by adding the indicator function of  $C$  to the first DC component.

Note that the subgradient of convex function  $\phi$  at  $x_0$  is defined by

$$\partial\phi(x_0) := \{y \in \mathbb{R}^n : \phi(x) \geq \phi(x_0) + \langle x - x_0, y \rangle, \forall x \in \mathbb{R}^n\}$$

The main principle of DCA is quite simple, that is, at each iteration of DCA, the convex function  $h$  is approximated by its affine minorant at  $y^k \in \partial h(x^k)$ , and it leads to solving the resulting convex program.

$$\begin{aligned} y^k &\in \partial h(x^k) \\ x^{k+1} &\in \arg \min_{x \in \mathbb{R}^n} \{g(x) - h(x^k) - \langle x - x^k, y^k \rangle\}. \end{aligned}$$

The computation of DCA is only dependent on DC components  $g$  and  $h$  but not the function  $f$  itself. Actually, there exist infinitely many DC decompositions corresponding to each DC function and they generate various versions of DCA. Choosing the appropriate DC decomposition plays a key role since it influences on the properties of DCA such as convergence speed, robustness, efficiency, globality of computed solutions, ... DCA is thus a philosophy rather than an



algorithm. For each problem we can design a family of DCA based algorithms. To the best of our knowledge, DCA is actually one of the rare algorithms for non-smooth nonconvex programming which allow to solve large-scale DC programs. DCA was successfully applied for solving various nonconvex optimization problems, which quite often gave global solutions and is proved to be more robust and more efficient than related standard methods [8–10, 13] and the list of reference in [3].

This is a DCA generic scheme:

- **Initialization.** Choose an initial point  $x^0$ ,  $0 \leftarrow k$ .
- **Repeat.**
  - Step 1.** For each  $k$ ,  $x^k$  is known, computing  $y^k \in \partial h(x^k)$ .
  - Step 2.** Calculating  $x^{k+1} \in \partial g^*(y^k)$ .  
where  $\partial g^*(y^k) = \arg \min_{x \in \mathbb{R}^n} \{g(x) - h(x^k) - \langle x - x^k, y^k \rangle : x \in C\}$
  - Step 3.**  $k \leftarrow k + 1$ .
- **Until** stopping condition is satisfied.

The convergence properties of DCA and its theoretical basis are analyzed and proved completely in [5, 8, 9].

The extension of DC programming and DCA was investigated for solving general DC programs with DC constraints [4] as follows:

$$\begin{aligned} & \min_x f_0(x), \\ & \text{s.t. } f_i(x) \leq 0, \forall i = 1, \dots, m, \\ & \quad x \in C, \end{aligned} \tag{1}$$

where  $C \in \mathbb{R}^n$  is a nonempty closed convex set;  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $\forall i = 0, 1, \dots, m$  are DC functions. It is apparent that this class of nonconvex programs is the most general in DC programming and as consequence it is more challenging to deal with than standard DC programs. In [4], the authors proposed two approaches for general DC programs to overcome the difficulty caused by the nonconvexity of the constraints. Both approaches are built on the main idea of the philosophy of DC programming and DCA, that is approximating (1) by a sequence of convex programs. The former was relied on penalty techniques in DC programming while the latter was based on the convex inner approximation method. In this paper, we use the idea of the second approach to solve the problem mentioned, the main scheme of this approach as follow:

Since  $m + 1$  functions  $f_i$  are DC functions, we can decompose them into difference of two convex functions.

$$f_i(x) = g_i(x) - h_i(x), x \in \mathbb{R}^n, i = 0, \dots, m.$$

By linearizing the concave parts of DC decompositions of all DC functions we derive sequential convex subproblems of the following form:

$$\begin{aligned} & \min_x g_0(x) - \langle y_0^k, x \rangle \\ & \text{s.t. } g_i(x) - h_i(x^k) - \langle y_i^k, x - x^k \rangle \leq 0, \forall i = 1, \dots, m \\ & \quad x \in C, \end{aligned} \tag{2}$$

where  $x^k \in \mathbb{R}^n$  is a point at the current iteration,  $y_i^k \in \partial h_i(x^k), \forall i = 0, \dots, m$ . The relaxation technique was proposed for problem (2) as following form

$$\begin{aligned} & \min_x \quad g_0(x) - \langle y_0^k, x \rangle + \lambda_k t & (3) \\ \text{s.t.} \quad & g_i(x) - h_i(x^k) - \langle y_i^k, x - x^k \rangle \leq t, \forall i = 1, \dots, m \\ & x \in C \\ & t \geq 0, \end{aligned}$$

where  $\lambda_k$  is a penalty parameter. The DCA scheme of the general DC program (3) as follows:

- **Initialization.** Choose an initial point  $x^0; \alpha_1, \alpha_2 \geq 0; \lambda_1 \geq 0$  and  $0 \leftarrow k$ .
- **Repeat.**
  - Step 1.** For each  $k, x^k$  is known, computing  $y^k \in \partial h_i(x^k), i = 0, \dots, m$ .
  - Step 2.** Calculating  $(x^{k+1}, t^{k+1})$  as the solution of (3), and the associated Lagrange multipliers  $(\beta^{k+1}, \mu^{k+1})$ .
  - Step 3.** Penalty parameter update.  
compute  $v_k = \min\{\|x^{k+1} - x^k\| - 1, \|\beta^{k+1}\|_1 + \alpha_1\}$
  - $\lambda_{k+1} = \begin{cases} \lambda_k & \text{if } \lambda_k \geq v_k \\ \lambda_k + \alpha_2 & \text{otherwise} \end{cases}$
  - Step 4.**  $k \leftarrow k + 1$ .
- **Until** stopping condition is satisfied.

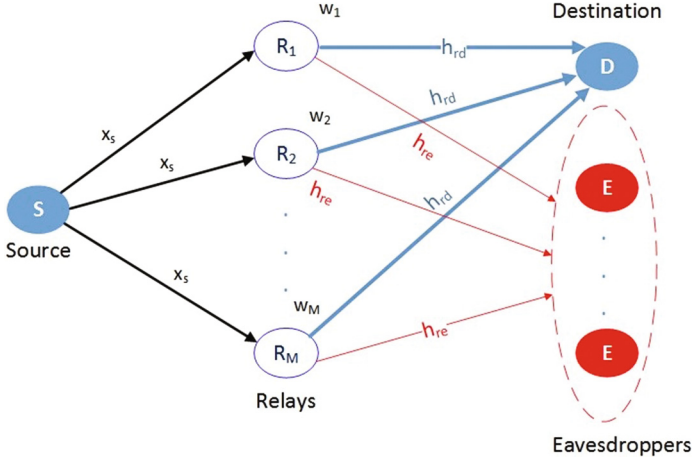
The global convergence of the general DC program above is shown in [4].

## 2 System Model and Received SNR Maximization Problem

### 2.1 System Model

In this paper, we consider the one-way communication system as in [11, 12]. The system consists of a single source ( $S$ ), a single destination ( $D$ ),  $M$  relays nodes ( $i \in \mathcal{M} = \{1, 2, \dots, M\}$ ) and  $K$  eavesdroppers ( $k \in \mathcal{K} = \{1, 2, \dots, K\}$ ) as shown in Fig. 1. The channel gain from a node  $p$  to a node  $q$  is denoted by a complex constant  $h_{p,q}$ , where  $p \in \{S\} \cup \mathcal{M}$  and  $q \in \{D\} \cup \mathcal{M} \cup \mathcal{K}$ . We assume the availability of complete channel state information (CSI), *i.e.*, all  $h_{p,q}$  are known throughout the network.

There are two hops in AF scheme. In the first hop, the source  $S$  transmit signal  $x_s$  to its trusted relays using the first transmission slot with power  $E(|x_s|^2) = P_s$ . The received signal at each relay node due to the source can be expressed as:  $y_i = h_{s,i}x_s + z_i$ ,  $i \in \mathcal{M}$ , where  $z_i$  is the background noise at  $i^{th}$  relay node that has a Gaussian distribution with zero mean and variance of  $\sigma^2$ . In the second hop, each relay node  $i$  scales the received signal from the source by  $\beta_i \in \mathbb{C}$  before transmitting to destination, the  $i^{th}$  relay node send the signal  $x_{r,i} = \beta_i y_i$  with the condition  $|\beta_i|^2 \leq \beta_{max,i}^2 = \frac{P_i}{|h_{s,i}|^2 P_s + \sigma^2}, \forall i \in \mathcal{M}$ .



**Fig. 1.** The wireless communication system model

The received signal at the destination and the eavesdroppers can be expressed as:

$$y_d = \sum_{i=1}^M h_{i,d} \beta_i y_i + z_d = \sum_{i=1}^M h_{s,i} \beta_i h_{i,d} x_s + \sum_{i=1}^M \beta_i h_{i,d} z_i + z_d, \quad (4)$$

$$y_l = \sum_{i=1}^M h_{i,l} \beta_i y_i + z_l = \sum_{i=1}^M h_{s,i} \beta_i h_{i,l} x_s + \sum_{i=1}^M \beta_i h_{i,l} z_i + z_l, \quad l = 1, 2, \dots, K, \quad (5)$$

where  $z_d$  and  $z_l$  are the complex white background Gaussian noise.

Then we can write the received SNR at the destination as:

$$SNR_d = \frac{|\sum_{i=1}^M h_{h,i} \beta_i h_{i,d}|^2 P_s}{1 + \sum_{i=1}^M |\beta_i h_{i,d}|^2 \sigma^2} \quad (6)$$

and the eavesdroppers as:

$$SNR_l = \frac{|\sum_{i=1}^M h_{h,i} \beta_i h_{i,l}|^2 P_s}{1 + \sum_{i=1}^M |\beta_i h_{i,l}|^2 \sigma^2}, \quad l = 1, 2, \dots, K. \quad (7)$$

## 2.2 Received SNR Maximization Problem

In this section, we consider the problem of maximizing the received SNR achievable with the AF-relays when the received SNR at the eavesdroppers is below their respective predefined thresholds. However, given the logarithmic dependence of the achievable information rate on the received SNR ( $I(P_s) \propto \log(1 + SNR)$ ) and the monotonically increasing nature of the  $\log(\cdot)$  function, we have:

$$\max_{\beta} \frac{|\sum_{i=1}^M h_{s,i} \beta_i h_{i,d}|^2 P_s}{1 + \sum_{i=1}^M |\beta_i h_{i,d}|^2 \sigma^2} \quad (8)$$

$$\text{such that: } \frac{|\sum_{i=1}^M h_{s,i} \beta_i h_{i,k}|^2 P_s}{1 + \sum_{i=1}^M |\beta_i h_{i,k}|^2 \sigma^2} \leq \gamma_k; \quad k \in \mathcal{K}$$

$$|\beta_i|^2 \leq \beta_{max,i}^2, \quad i \in \mathcal{M}.$$

Here,  $\gamma_k$  is a real number and represents the predefined threshold for the  $k^{th}$  eavesdropper. This problem is inherently non-convex and hard to solve. We consider the following transformation of the variable  $\beta_i$ .

$$w_i = \beta_i h_{i,d} \text{ and } u_i = \frac{w_i}{\sqrt{1 + w^\dagger w}}$$

If we consider the vector variables  $\mathbf{u} = [u_1, u_2, \dots, u_M]^T$  and  $\mathbf{w} = [w_1, w_2, \dots, w_M]^T$ , then we can write:

$$\mathbf{u} = \frac{\mathbf{w}}{\sqrt{\mathbf{1} + \mathbf{w}^\dagger \mathbf{w}}} \Leftrightarrow \mathbf{w} = \frac{\mathbf{u}}{\sqrt{\mathbf{1} - \mathbf{u}^\dagger \mathbf{u}}}$$

Also we define  $\rho_{\mathbf{k}} = [\rho_{1,k}, \dots, \rho_{M,k}]$  where  $\rho_{i,k} = \frac{|h_{i,k}|}{|h_{i,d}|}$ .

In terms of these new variables and parameters, the problem (8) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{u}} \quad & -\mathbf{u}^\dagger \mathbf{h}_s \mathbf{h}_s^\dagger \mathbf{u} \\ \text{s.t.} \quad & \mathbf{u}^\dagger \mathbf{C}_k \mathbf{u} \leq 1, \quad k \in \mathcal{K} \\ & \mathbf{u}^\dagger \mathbf{D}_i \mathbf{u} \leq 1, \quad i \in \mathcal{M}, \end{aligned} \quad (9)$$

where  $\mathbf{h}_s = [h_{s,1}, \dots, h_{s,M}]^\dagger$ ,  $\mathbf{D}_{\rho,k} = \text{diag}(|\rho_k|^2)$ ,  $\gamma'_k = \gamma_k \frac{\sigma^2}{P_s}$ ,  $\forall k \in \mathcal{K}$ ;

$$\mathbf{C}_k = \frac{\mathbf{h}_{s\rho,k} \mathbf{h}_{s\rho,k}^\dagger}{\gamma'_k} + \mathbf{I} - \mathbf{D}_{\rho,k} \text{ where } \mathbf{h}_{s\rho,k} = [h_{s,1}\rho_{1,k}, h_{s,2}\rho_{2,k}, \dots, h_{s,M}\rho_{M,k}]^\dagger$$

$$\text{and } (\mathbf{D}_i)_{jk} = \begin{cases} 1 + \frac{1}{|h_{i,d}|^2 \beta_{i,max}^2}, & \text{if } k = j = i \\ 1, & \text{if } k = j \neq i \\ 0, & \text{otherwise.} \end{cases}$$

As the objective function is nonconvex and the constraints could be convex or not, if  $\rho_{i,k} = \frac{|h_{i,k}|}{|h_{i,d}|} \leq 1, \forall i, k$  then  $\mathbf{I} - \mathbf{D}_{\rho,k}$  is diagonal matrix with positive entries, therefore,  $\mathbf{C}_k$  is a positive definite matrix so all the constraints are convex. But, in general scenarios,  $\mathbf{C}_k$  may not be positive-semidefinite then the first constraint is nonconvex. Therefore, the problem (9) is hard to get the optimal solution in general.

*Remark:* The AF secrecy rate of the system would be  $\frac{1}{2} \log(1 + SNR)$ , where  $SNR$  is the optimum objective function value of problem (9).

### 3 Existing work

The problem (9) has form of *Quadratically Constrained Quadratic Program* (QCQP) with nonconvex objective function and nonconvex constraint. It is

difficult to find the optimal solution of that problem by solving directly in general. The existing method proposed in [11, 12] is to find suboptimal solution by *Semi-definite Relaxation* (SDR) method.

By define  $\mathbf{U} = \mathbf{u}\mathbf{u}^\dagger$ , and considering relaxation on rank one symmetric positive semi-definite (PSD) constraint ( $\text{rank}(\mathbf{U}) = 1$ ), the optimization program (9) can be written as:

$$\begin{aligned} & \max_{\mathbf{U}} \text{trace}(\mathbf{h}_s \mathbf{h}_s^\dagger * \mathbf{U}) \\ & \text{subject to: } \text{trace}(\mathbf{C}_k * \mathbf{U}) \leq 1, k \in \mathcal{K} \\ & \text{trace}(\mathbf{D}_i * \mathbf{U}) \leq 1, i \in \mathcal{M}. \end{aligned} \quad (10)$$

As the objective and all constraints in (10) are convex, this problem can be solve by CVX optimization tool. Once the problem (10) is solved, we can find the corresponding optimal  $\mathbf{u}$  and thereby  $\mathbf{w}$  by applying eigenvalue decomposition on matrix  $\mathbf{U}$ .

#### 4 DC Programming and DCA for Problem (9)

In this section, we propose a DC decomposition then derive a DCA scheme for the original received SNR maximization at the destination problem (9). By define

$$\begin{aligned} \rho_k^+ &= \begin{cases} 1 - |\rho_{i,k}|^2, & \text{if } |\rho_{i,k}| \leq 1 \\ 0, & \text{else} \end{cases} \\ \rho_k^- &= \begin{cases} |\rho_{i,k}|^2 - 1, & \text{if } |\rho_{i,k}| \geq 1 \\ 0, & \text{else} \end{cases} \end{aligned}$$

The problem (9) can be rewritten as

$$\begin{aligned} & \min_{\mathbf{u}} 0 - \mathbf{u}^\dagger \mathbf{H}_s \mathbf{u} \\ & \text{s.t. } \mathbf{u}^\dagger \mathbf{C}_k^+ \mathbf{u} - \mathbf{u}^\dagger \mathbf{C}_k^- \mathbf{u} \leq 1, \forall k \in \mathcal{K} \\ & \mathbf{u}^\dagger \mathbf{D}_i \mathbf{u} \leq 1, \forall i \in \mathcal{M}, \end{aligned} \quad (11)$$

where  $\mathbf{H}_s = \mathbf{h}_s \mathbf{h}_s^\dagger$ ;  $\mathbf{C}_k^+ = \frac{h_{s\rho,k} h_{s\rho,k}^\dagger}{\gamma_k} + \text{diag}(\rho_k^+)$  and  $\mathbf{C}_k^- = \text{diag}(\rho_k^-)$ .

The problem (11) above is actually a general DC program at the objective function and the first  $K$  constraints as the form:

$$\begin{aligned} & \min_{\mathbf{u}} G_1(\mathbf{u}) - H_1(\mathbf{u}) \\ & \text{s.t. } G_2^k(\mathbf{u}) - H_2^k(\mathbf{u}) \leq 1, \forall k \in \mathcal{K} \\ & \mathbf{u}^\dagger \mathbf{D}_i \mathbf{u} \leq 1, \forall i \in \mathcal{M}, \end{aligned} \quad (12)$$

where  $G_1(\mathbf{u}) = 0$ ;  $H_1(\mathbf{u}) = \mathbf{u}^\dagger \mathbf{H}_s \mathbf{u}$ ;  $G_2^k(\mathbf{u}) = \mathbf{u}^\dagger \mathbf{C}_k^+ \mathbf{u}$ ; and  $H_2^k(\mathbf{u}) = \mathbf{u}^\dagger \mathbf{C}_k^- \mathbf{u}$ .

With the  $K + 1$  functions  $H_1(\mathbf{u})$  and  $H_2^k(\mathbf{u})$  in (12) are convex and smooth, we apply DCA to solve this problem as the following DCA scheme.

**DCA Scheme:****Initialization:** Choose a random initial point  $u^0$ ,  $\lambda > 0$ ,  $l = 0$ **Repeat:**  $l = l + 1$ , Calculating  $u^l$  by solving this subproblem:

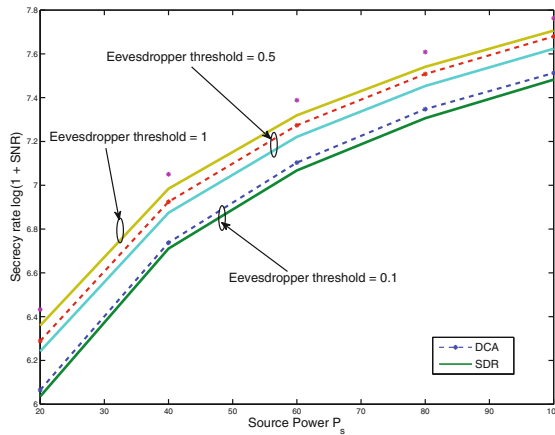
$$\begin{aligned} \min_{\mathbf{u}, t} \quad & -\mathbf{u}^\dagger (\mathbf{H}_s \mathbf{u}^{l-1}) - (\mathbf{H}_s \mathbf{u}^{l-1})^\dagger \mathbf{u} + \lambda t \quad (13) \\ \text{s.t.} \quad & \mathbf{u}^\dagger \mathbf{C}_k^+ \mathbf{u} - \left[ \mathbf{u}^\dagger (\mathbf{C}_k^- \mathbf{u}^{l-1}) + (\mathbf{C}_k^- \mathbf{u}^{l-1})^\dagger \mathbf{u} \right] \leq 1 + (\mathbf{u}^{l-1})^\dagger \mathbf{C}_k^- \mathbf{u}^{l-1} + t, \forall k \in \mathcal{K} \\ & \mathbf{u}^\dagger \mathbf{D}_i \mathbf{u} \leq 1, \forall i \in \mathcal{M} \\ & t \geq 0. \end{aligned}$$

**Until:**  $\left( \frac{\|\mathbf{u}^l - \mathbf{u}^{l-1}\|}{1 + \|\mathbf{u}^{l-1}\|} \leq \epsilon \text{ or } \frac{|f(\mathbf{u}^l) - f(\mathbf{u}^{l-1})|}{1 + |f(\mathbf{u}^{l-1})|} \leq \epsilon \right)$ , where  $f(\mathbf{u}^l) = (\mathbf{u}^l)^\dagger \mathbf{H}_s \mathbf{u}^l$ .

## 5 Numerical Results

In this section, we study the performance of the proposed method and compare with its of the existing SDR scheme which gave a suboptimal solution [11, 12]. For simplicity, we take into account a simple one-dimensional communication system model, as illustrated in Fig. 1, mentioned in Sect. 2. In our experiment, we test with the number of relays  $M = 15$  with respect to number of eavesdroppers  $K = 5$ . The source power varies from 20 to 100, and three levels of predefined eavesdropper thresholds are  $\gamma = 0.1, 0.5$  and  $1$ . We perform experiments consisting of 100 independent trials to obtain the average results. For each trial, the channel gains are obtained from a complex Gaussian distribution with zero mean and variance  $\sigma^2 = 1$ .

This is apparent from Fig. 2 that, all the secrecy rates are rising when the source power consumption ( $P_s$ ) increases. In all three cases of the predefined



**Fig. 2.** Secrecy rate vs. Power relay constraint; Number of Relays  $M = 15$ , number of eavesdroppers  $K = 5$

thresholds, DCA (dash line) always provides the better values of secrecy rate than SDR (solid line) does. In particular, it shows that the gap of them increases with the value of eavesdropper thresholds ( $\gamma = 0.1, 0.5, 1.0$ ).

## 6 Conclusion

In this paper, we have investigated physical layer security in the Amplify-and-Forward relay beamforming networks in the presence of multiple eavesdroppers based on the SNR approach. After motivating the secure communication scheme, we formulated the received SNR maximization problem as a DC programming and applied DCA scheme to solve it. The numerical results show the efficiency of our approach compared with that of the existing method.

The proposal for applying DC programming and DCA to deal with SNR based approach on other wireless cooperative networks such as the Decode-and-Forward scheme or cooperative jamming scheme are being studied in our future works.

**Acknowledgements.** We would like to thank Prof. Le Thi Hoai An, and the DCA research group in LITA, University of Lorraine, France for helping us do research in this field. We also thank the referees for leading us to improve the manuscript.

## References

1. Csiszár, I., Körner, J.: Broadcast channels with confidential messages. *IEEE Trans. Inf. Theory* **24**(3), 339–348 (1978)
2. Dong, L., Han, Z., Petropulu, A., Poor, H.: Improving wireless physical layer security via cooperating relays. *IEEE Trans. Sig. Process.* **58**(3), 1875–1888 (2010)
3. Le Thi, H.A.: DC Programming and DCA. <http://www.lita.univ-lorraine.fr/~lethi/>
4. Le Thi, H.A., Huynh, V.N., Pham, D.T.: DC programming and DCA for general DC programs, vol. 282. Springer (2014)
5. Le Thi, H.A., Pham, D.T.: The DC (Difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.* **133**, 23–46 (2005)
6. Leung-Yan-Cheong, S.K., Hellman, M.E.: The Gaussian wire-tap channel IT-24, 451–456, July 1978
7. Dong, L., Han, Z., Petropulu, A.P., Poor, H.V.: Improving wireless physical layer security VI cooperating relays. *IEEE Trans. Sig. Process.* **58**(5), 1875–1888 (2010)
8. Pham, D.T., Le Thi, H.A.: Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Math. Vietnamica* **22**(1), 289–357 (1997)
9. Pham, D.T., Le Thi, H.A.: Optimization algorithms for solving the trust region subproblem. *SIAM J. Optim.* **8**, 476–505 (1998)
10. Pham, D.T., Le Thi, H.A.: Recent Advances in DC Programming and DCA, vol. 8342. Springer, Heidelberg (2014)
11. Siddhartha Sarma, S.A., Kuri, J.: Secure communication in Amplify-and-Forward networks with multiple eavesdroppers: decoding with SNR thresholds. *Wireless Pers. Commun.* **85**(4), 1945–1956 (2015)

12. Siddhartha Sarma, S.A., Kuri, J.: Beam-forming for secure communication in Amplify-and-Forward Networks: An SNR based approach, May 2014. [arXiv:1406.4844](https://arxiv.org/abs/1406.4844)
13. Tran, T.T., Le Thi, H.A., Pham, D.T.: DC programming and DCA for a novel resource allocation problem in emerging area of cooperative physical layer security. In: Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing. Springer, Cham **358**, 57–68 (2015)
14. Wyner, A.D.: The wire-tap channel. Bell Sys. Tech. J. **54**, 1355–1387 (1975)
15. Yang, Y., Li, Q., Ma, W.K.: Cooperative secure beamforming for AF relay networks with multiple eavesdroppers. IEEE Sig. Process. Lett. **20**(1), 35–38 (2013)



# A Cash-Flow-Based Optimization Model for Corporate Cash Management: A Monte-Carlo Simulation Approach

Linus Krumrey<sup>1</sup>, Mahdi Moeini<sup>2</sup>(✉), and Oliver Wendt<sup>2</sup>

<sup>1</sup> Treasury Intelligence Solutions, Alte Schönhauserstr. 44, 10119 Berlin, Germany  
linus.krumrey@tis.biz

<sup>2</sup> BISOOR, Technical University of Kaiserslautern, Postfach 3049,  
Erwin-Schrödinger-Str., 67653 Kaiserslautern, Germany  
{mahdi.moeini,wendt}@wiwi.uni-kl.de

**Abstract.** *Cash management* is an essential task to ensure a company's ongoing success, and consists in ensuring the company's solvency, minimizing risks, and maximizing the financial result. In this paper, we propose a new cash-flow-based model for cash management. This model considers the risk of customer default, the risk of bank default, and the foreign exchange risk, and aggregates them into a single *overall risk*. Finally, we report the results of numerical experiments, carried out on a case study, via Monte-Carlo simulation, and by using our bi-objective optimization model. The results confirms the validity of the introduced model.

**Keywords:** Cash management · Risk management · Cash Flow at Risk · Monte-Carlo simulation

## 1 Introduction

In financial literature, the term *cash management* refers to all measures undertaken for short term financial planning [11]. More precisely, cash management is defined as managing cash flows in order to ensure solvency, to minimize risk, and to maximize financial result [11, 16].

### 1.1 Literature Review

Cash management models have already been investigated in the literature. A first cash management model, the Baumol-Model, in which inventory management approaches are applied to stocks of cash, was introduced in [2]. It assumes that all costs, interests, and transactions are constant and known in advance. As an extension of the Baumol-Model, [13] introduced the Miller-Orr-Model, which models the cash development of firms. In the Miller-Orr-Model, the in- and out-flow of cash is modeled as a stationary random walk that, in every time interval, generates either an in- or an outflow of a fixed size [13]. The Miller-Orr-Model

has been extended numerous times (see e.g., [7, 10]). As criticized in [8], the Miller-Orr-Model considers both in- and outflow of cash to be random. However, in a *Business-to-Business* context, in- and outflow are usually not random but result from contracts [16]. As an alternative approach, network flow models were introduced by Golden et al. [6]. In network flow models, the vertices represent the bank accounts of a firm and the edges show the cash flows. The objective is to maximize the flow into the “goal node”, i.e., a vertex with zero outgoing edges. One major strength of network flow models is that they can be transformed into linear programs and solved efficiently [6]. A recent extension of this model is presented in [1], where the authors applied the model from [6] to a Brazilian company and extend the model to also include tactical cash management, namely the planning of loan payments. The network flow models assume that all cash flows are completely deterministic [6]; however, this does not always reflect reality as incoming cash flows are in fact *not* completely deterministic. Both monetary value and the arrival time carry risk. More precisely, the actual amount of an incoming cash flow may be less than defined in the contract [9]. Additionally, incoming cash flows may be delayed and arrive after the originally expected arrival date [12].

Both approaches, the stochastic ones based on the Miller-Orr-Model and the network flow models, do not differ between inflow and outflow. In stochastic models, both in- and outflow are random and in network flow models, in- and outflow are deterministic. However, in reality, outgoing cash flows (i.e., date of payment and amount of each cash flow) are decision variables of the firm [16].

## 1.2 Contributions of this Paper

In this paper, we propose a new simulation-based cash management model with the purpose of removing some of the simplifications implied in the previous models. Indeed, our model assumes incoming cash flows with expected arrival date and payout; however, both the arrival date and the payout carry risk, i.e., incoming cash flows may arrive too late and the payout may be less. Outgoing cash flows also have an expected execution date and monetary value, but both the execution date and the payout are decision variables of the firm. As an extension of [11], we consider not only the foreign exchange risk but also the customer risk and the banking risk. Furthermore, for both incoming and outgoing cash flows, the firm may choose among all of its different bank accounts, that may be located at different banks and in different countries [16]. This is a fundamental aspect, since interest rates, transaction costs, and transaction durations may vary from country to country, from bank to bank, and even from account to account.

The paper is organized as follows: In Sect. 2, we introduce the main notation and basic concepts that we need in this paper. In Sect. 3, we present the objective functions and we introduce the different types of risk considered by this model. In Sect. 4, we explain how our simulation algorithm calculates and quantifies the risks. Afterwards, in Sect. 5, we examine a short case study using our model. We finish this paper with concluding remarks in Sect. 6.

## 2 Notation and Basic Definitions

Let us introduce the notations that we are going to use throughout this paper. Let  $t_0$  be the first and  $t_n$  the last period of the planning horizon and assume that  $c \in C$  represents a single currency in the set of currencies  $C$ . The base currency, i.e., the main currency of the firm under observation, is denoted by  $c_b$ . The fraction  $(c_1/c_2)_t$  denotes the exchange rate from  $c_1$  to  $c_2$  in period  $t$ . Further,  $A$  denotes the set of all bank accounts and  $a \in A$  is a single account. We denote by  $ICF_{a,t}^*$  the set of all *incoming* cash flows that arrive at bank account  $a$  and that are due in period  $t^*$ . Furthermore, the set  $ICF_{a,t}$ , generated via Monte-Carlo simulation [14], contains the same cash flows as  $ICF_{a,t}^*$  but with a less optimistic estimate for the arrival times of the cash flows (for more details, see also [12]). The set of all *outgoing* cash flows that are paid from bank account  $a$  in period  $t$  is denoted by  $OCF_{a,t}$ . Additionally, let  $icf_{a,t} \in ICF_{a,t}$  denote a single incoming and  $ocf_{a,t} \in OCF_{a,t}$  a single outgoing cash flow. For a given set of all business partners  $P$ , the function  $r : \{ICF_{a,t} | \forall a \in A, t_0 \leq t \leq t_n\} \mapsto P$  returns the origin of  $icf$ . The function  $\bar{c} : \{\{ICF_{a,t} \cup OCF_{a,t} | \forall a \in A, t_0 \leq t \leq t_n\} \cup A\} \mapsto C$  maps cash flows to the currency in which they are paid and bank accounts to the currency in which they are accounted. The cash balance on bank account  $a$  in the *beginning* of period  $t$  is denoted by  $L_{a,t}$  (which is equals to the cash balance in the *end* of period  $t - 1$ ). Also,  $l_{a,t}$  denotes the credit line available for bank account  $a$  in period  $t$ . Thus, the available *liquidity* on a bank account  $a$  in period  $t$  is  $L_{a,t} + l_{a,t}$ . The cash balance is computed recursively for each  $t$  by

$$L_{a,t} = L_{a,t-1} + \sum_{icf \in ICF_{a,t-1}} icf - \sum_{ocf \in OCF_{a,t-1}} ocf. \quad (1)$$

The recursion ends as soon as  $t$  reaches  $t_{-1}$  (current period) with known cash balance.

For the risk simulation, we use  $\alpha$  to indicate the confidence level and we suppose that  $min_{icf}$ ,  $mode_{icf}$ , and  $max_{icf}$  denote the minimal, the most likely, and the maximal loss for an incoming cash flow  $icf$  at loss. Finally, we define  $B$  as the set containing all relevant banks  $b \in B$ . More specifically, the function  $\bar{b} : A \mapsto B$  maps an *account* to its governing *bank*. We define the function  $pd : \{\{ICF_{a,t} \cup OCF_{a,t} | \forall a \in A, t_0 \leq t \leq t_n\} \cup B \cup P\} \mapsto [0, 1[$  to return the *probability of default* of an incoming cash flow, a customer, or a bank. Finally, the function  $g(b)$  returns the country in which a given bank  $b$  is located.

## 3 The Cash Management Model

### 3.1 Objective Functions and Constraints

In this section, we describe the objective functions and the constraints of our mathematical model. First, we note that a vital part of cash managing is *ensuring*

*solvency*, i.e., the firm must always have enough liquidity at hand to execute all planned outgoing cash flows on time:

$$\sum_{ocf \in OCF_{a,t}} ocf \leq L_{a,t} + l_{a,t} \quad : \quad \forall a \in A, t_0 \leq t \leq t_n. \quad (2)$$

Constraints (2) enforce that the expected liquidity in each period  $t$  is sufficient to pay for all outgoing cash flows in period  $t$ . Since one cannot rely on incoming cash flows to be always on time [12], the incoming cash flows expected in period  $t$  are not used in these constraints.

Second, the *financial result* is that part of a firm's profit that is not induced by its normal business activity, but comes from activities on the financial market. It can be increased by increasing the credit interest, decreasing the debit interest, and/or by reducing transaction costs. Due to the fact that cash managers are constrained in their actions, maximizing the financial result is equivalent to maximizing the liquidity at the end of the planning horizon [16]. We first define the function  $\Gamma(CF, c, t)$  that calculates the sum of all cash flows in a set of cash flows  $CF$  and converts each of them into the currency  $c$ , using the exchange rate from period  $t$ :

$$\Gamma(CF, c, t) = \sum_{cf \in CF} \left( \frac{c}{\bar{c}(cf)} \right)_t \cdot cf. \quad (3)$$

Let  $F_{c_b, t_0, t_n}$  denote the total expected case cash balance change converted into the base currency  $c_b$  at the end of period  $t_n$ . The first objective function is:

$$\max F_{c_b, t_0, t_n} = \sum_{a \in A} \left( \sum_{t=t_0}^{t_n} \Gamma(ICF_{a,t}, c_b, t_0) - \Gamma(OCF_{a,t}, c_b, t_0) \right). \quad (4)$$

The result of  $\Gamma(ICF_{a,t}, c_b, t_0) - \Gamma(OCF_{a,t}, c_b, t_0)$  is the total net cash flow for bank account  $a$  in period  $t$  converted to the base currency  $c_b$ . The total net cash flows from all periods are summed up for each bank account  $a$ . Note that there is no interest calculation or discounting of the in- and outflow in this formula. In practice, interest is calculated daily, but only paid monthly, quarterly, or yearly [16]. For simplicity, we assume that the interest payment is not within the planning horizon and we can thus, in this paper, disregard this topic.

The second objective function is to minimize the difference  $R$  between the risk  $CFaR$  (see Sect. 4) and a *target risk*  $tr$ . The target risk is a non-negative value and represents a risk level that a firm desires to reach (in practice, the target risk is often considered to be equal to 0).

$$\min R = |CFaR - tr|. \quad (5)$$

### 3.2 Types of Risk in Cash Flows

The International Organization for Standardization (ISO) defines the concept of *risk* as the effect of uncertainty on objectives. The *effect* includes positive and

negative deviations from the plan. *Financial risk* translates this concept into the world of finance and focuses on the negative deviation. In [3], financial risk is defined as a loss that occurs as a result of disadvantageous price or rate changes or disadvantageous changes on the financial market. We focus this definition of risk on *cash flows* and therefore interpret *risk* as the existence of an unexpected outflow or the absence of an expected inflow. Hence, the risk can be quantified as the difference between the *planned case* and the *expected worst case* (see also Sect. 4). In this paper, the focus lies on four types of risk:

- **The Risk of Default of Customers:** We count a customer as *defaulted* when he is either not able or not willing to pay his liabilities (partially or completely).
- **The Risk of Default of Banks:** The financial crisis, which started 2007 and peaked with the bankruptcy of Lehman Brothers in 2008, has moved the risk associated with banks into the focus of risk managers [4]. If a bank becomes insolvent, all cash on bank accounts which exceeds the deposit insurance in that specific country is generally lost [5].
- **The Foreign Exchange Risk:** When a firm conducts business in more than one currency, then there is the risk that the exchange rate between these currencies changes in an unfavorable way. This risk is called *foreign exchange (FX) risk*. According to [15], there are generally three components of the foreign exchange risk: economic risk, translation risk, and the transaction risk. *Economic risk* and *translation risk* describe the influence of exchange rates on a firm's balance sheet and business model respectively. In this paper, we focus on the *transaction risk*, which stems from accounted but not yet paid or received invoices. Between the planning and the actual payment, the exchange rate may fluctuate, leading to a difference between the accounted amount and the actually paid/received amount [15].
- **The Overall Risk:** For most firms, the three types of financial risk presented above do not arise isolated from each other. For example, fluctuations of the foreign exchange rates have a direct influence on the incoming cash flows from customers. Hence, an encompassing view on the financial risk, called *overall risk*, is in a firm's interest.

## 4 Cash Flow at Risk

### 4.1 Description of the Cash Flow at Risk

The *Value at Risk* (VaR) is a common financial risk measure in practice [9]. The VaR-methodology implicitly assumes that the structure of the portfolio does not change within the holding period. While this assumption is acceptable for financial assets, it does not hold for cash flows since cash flows are dynamic in nature. Hence, the VaR concept is not applicable to cash flows. However, there is another concept called the *Cash Flow at Risk* (CFaR), presented in [9]. It embraces the dynamic nature of cash flows and results in a single value that is interpreted similarly to the VaR. Indeed, under normal market conditions, with

the given confidence level, at the end of the observation period, the difference between the *planned* and the *actual* inflow will be at most the CFaR.

The CFaR is determined via a *Monte-Carlo simulation* [9]. Monte-Carlo simulations are a very general class of procedures which are used in many different research areas. More precisely, Monte-Carlo simulations try to predict the future by creating a multitude of random scenarios. Analyzing all generated scenarios allows good predictions for the future. Monte-Carlo simulations are known to be powerful and flexible, albeit they may be computationally very expensive [14].

We use Algorithm 1 in Sect. 4.2 to calculate the CFaR (i.e., *overall risk*).

## 4.2 Determining the Overall Risk

Before determining the overall risk, we need to define some functions. First, we present the following function in order to simulate exchange rate changes:

$$\begin{pmatrix} c_j \\ c_k \end{pmatrix}_t = (1 + rn_{t-1}) \begin{pmatrix} c_j \\ c_k \end{pmatrix}_{t-1}, \quad (6)$$

where,  $rn_{t-1} \in [-1, +\infty[$  is drawn at random and represents the relative change of the exchange rate from  $t - 1$  to  $t$ . The probability distribution, from which  $rn_{t-1}$  is drawn, can be fitted to sample data for the exchange rates or chosen by an expert.

Further, we need a function to filter sets of cash flows for those in a specific currency. For this purpose, the following function returns the cash flow  $cf$  if and only if the cash flow's currency matches the provided currency  $c$ .

$$\omega(c, cf) = \begin{cases} cf & : \text{if } c = \bar{c}(cf), \\ 0 & : \text{otherwise.} \end{cases} \quad (7)$$

We use the function (7) to define the following function with the objective of calculating the sum of all cash flows in a set of cash flows  $CF$  in currency  $c$ :

$$\phi(CF, c) = \sum_{cf \in CF} \omega(c, cf). \quad (8)$$

Last but not least, we define two *damage functions*, one for *incoming cash flows* and the other one for *banks*. Function (9) defines the damage function for incoming cash flows, where we need to distinguish between two cases: Either the origin of the cash flow is bankrupt and thus cannot pay at all, or the cash flow is not paid entirely for other reasons. In the first case, the damage is the complete cash flow. In the latter case, the damage is drawn randomly from a triangular distribution  $\Delta$  which is parameterized with  $min_{icf}$ ,  $mode_{icf}$ , and  $max_{icf}$ .

$$dm(icf) = \begin{cases} icf & : \text{if } r(icf) \text{ defaulted,} \\ \sim \Delta (min_{icf}, mode_{icf}, max_{icf}) & : \text{otherwise.} \end{cases} \quad (9)$$

Function (10) computes the damage for a single bank  $b$  that went bankrupt in period  $t$ .

$$dm(b, t) = \sum_{a \in \{A | \bar{b}(a) = b\}} \left( \frac{c_b}{\bar{c}(a)} \right)_t \cdot max(L_{a,t}, 0). \quad (10)$$

---

**Algorithm 1.** An algorithm to calculate the overall risk per period.

---

```

1 Inputs:  $\Sigma$ : a set of all  $\Sigma_c$ 
2            $DI$ : set containing the deposit insurance per country
3            $C$ : set of all currencies
4            $B$ : set of all banks
5            $A$ : set of all bank accounts
6            $ICF$ : set of all incoming cash flows
7            $OCF$ : set of all outgoing cash flows
8 Output:  $OR$ : set of overall risks per period
9
10 Start (Computing the Overall Risk per Period):
11    $OR = \{\}, BB = \{\}$ 
12   for  $t = t_0$  to  $t_n$  do:
13     compute  $(c^b/c_i)_t$  for all currencies  $c_i$  via (6)
14     for each  $p : P$  do:
15       if  $pd(p) \geq$  random number  $\in [0; 1]$ :
16         set flag that origin defaulted to true
17     end
18     for each  $c_1 \in C$  do:
19        $l = 0$ 
20       for each  $c_2 \in C$  do:
21          $e = s = 0$ 
22         for each  $a \in A$  where  $\bar{c}(a) = c_1$  do:
23            $e += \phi(ICF_{a,t}, c_2) - \phi(OCF_{a,t}, c_2)$  via (8)
24            $s += \phi(ICF_{a,t}, c_2) - \phi(OCF_{a,t}, c_2)$  via (8)
25           for each  $icf_{a,t} \in ICF_{a,t}$  do:
26             if  $(\bar{c}(icf_{a,t}) == c_2)$ :
27               if (origin defaulted | |
27                  $pd(icf_{a,t}) \geq$  random number  $\in [0; 1]$ ):
28                  $s -= dm(icf_{a,t})$  via (9)
29             end
30           end
31            $l += (c^1/c_2)_0 \cdot e - (c^1/c_2)_t \cdot s$ 
32         end
33        $OR_t += (c^b/c_1)_t \cdot l$ 
34     end
35     for each  $b \in B$  do:
36       if  $(b \notin BB \ \&\& \ pd(b) \geq$  random number  $\in [0; 1])$ :
37          $BB_t \cup = b$ 
38          $OR_t += (c^b/(\bar{c}(b)))_t \cdot \max(dm(b, t) - DI_{g(b)}, 0)$  via (10)
39       end
40      $OR = OR \cup \{OR_t\}$ 
41   end
42   return  $OR$ 
43 End

```

---

Algorithm 1 determines the overall risk (i.e., CFaR). In fact, the simulation starts by iterating over every period. The first step (line 13) consists in simulating the exchange rates for the current period. Afterwards, we iterate over every currency  $c_1$ , where the local variable  $l$  represents the loss in  $c_1$ . The currency  $c_1$  is used to iterate over all bank accounts  $a$  accounted in  $c_1$  and to collect the losses per currency across all bank accounts, while currency  $c_2$  is used to calculate the losses for each *individual bank account* in each currency. The variable  $e$  stores the *expected* net flow in  $c_2$  while  $s$  stores the *simulated* net flow in  $c_2$  (lines 21 and 23). After adding the expected net flow in  $c_2$  to  $s$  (line 24), we subtract all the simulated losses from  $s$  (lines 25 to 29). The loss is the difference between  $e$  and  $s$ , both converted to  $c_1$  using the appropriate exchange rate (line 31). The expected net flow  $e$  is converted with the known exchange rate  $(c_1/c_2)_{t-1}$ , while  $s$  is converted using the simulated exchange rate  $(c_1/c_2)_t$ . In order to calculate the correct overall risk  $OR_t$  for period  $t$ , we need to convert  $l$  from  $c_1$  into the base currency  $c_b$  and add it to  $OR_t$  (line 33). The incorporation of the risk of bank defaults starts in line 35. The algorithm iterates over every period and checks for every bank  $b$  whether it has already defaulted (in a previous period). If it did, no further action has to be done and we commence with the next bank. If the bank did not already default, the algorithm determines whether it defaults in the current period by comparing its probability of default  $pd(b)$  with a random number  $\in [0, 1]$ . If the random number is smaller than  $pd(b)$ ,  $b$  counts as defaulted. In this case, the damage is the total amount of cash at that bank subtracted by the corresponding deposit insurance, but at least 0. The damage is then converted to  $c_b$  and added to  $OR_t$ . The resulting series of values must be summed up in order to form a single value, the *scenario loss*  $SLO_{t_0, t_n}$ .

$$SLO_{t_0, t_n} = \sum_{t=t_0}^{t_n} OR_t, \quad (11)$$

where  $OR_t$  is the result of the simulation presented in Algorithm 1. The overall risk (i.e., the CFaR for our problem) is the  $(1 - \alpha)$  quantile of all scenarios.

## 5 Numerical Experiments

In order to validate the model, we conducted a case study on a *fictitious* company. For this purpose, we calculated, for this company, the presented four types of risk. We also checked which bank accounts violated the constraints (2). In period  $t_0$ , the risk is always EUR0 and we observe its development over time. We ran each simulation for “10000 · (number of *risk items*)” iterations, where a *risk item* is anything that may have some risk associated with it, i.e., a bank, an incoming cash flow, or a currency. Since the number of risk items is different for different risk types, the number of iterations varies accordingly.

### 5.1 Test Data

We assume that the planning horizon starts at  $t_0 = 1$  and terminates at  $t_n = 30$ . We compute the risks with the confidence level of  $\alpha = 95\%$ .



**Currencies:** We suppose that the company does business in four different currencies<sup>1</sup>: Euro (which is the base currency and thus has an exchange rate of 1 ( $^{\text{EUR}}/\text{EUR}$ )), US-Dollar (with an exchange rate of 0.88992 ( $^{\text{EUR}}/\text{USD}$ )), British Pound (1.18228 ( $^{\text{EUR}}/\text{GBP}$ )), and Swiss Franc (0.91341 ( $^{\text{EUR}}/\text{CHF}$ )).

**Banks and Bank Accounts:** In our case study, we suppose that the company uses three different banks in three countries, namely  $B1$ ,  $B2$ , and  $B3$  in Germany, the United States of America (USA), and the United Kingdom (UK), respectively. The probability of default for each of these banks is, respectively, 2%, 4%, and 3%. The deposit insurances are EUR 100000 in Germany, USD 250000 in the USA, and GBP 85000 in the UK. Furthermore, we assume that the company has six different bank accounts. All bank accounts have a credit line  $l_{a,t}$  of 0 in the complete planning horizon. For more details, see Table 1.

**Table 1.** The bank accounts used in the case study.

Bank	B1				B2	B3
Account	B1-1	B1-2	B1-3	B1-4	B2-1	B3-1
Initial balance	EUR 6000	EUR 0	EUR 17000	EUR 30000	USD 7000	GBP 5000

Without loss of generality, we assumed that there is no transaction cost. This assumption does not deteriorate the quality of this case study. Since the transaction costs apply in the *planned case* in the same way as in the *worst case*, the difference between the planned and the expected worst case remains identical. Thus, the *risk* is unaffected by transaction costs.

**Business Partners:** Additional to the banks and bank accounts, we assumed that the company has 12 *business partners*. Six of them are customers and the others are suppliers. We present the details for the customers in Table 2. In this table, the column *Currency* gives the currency in which the customer pays his cash flows and the column *Bank Account* shows the bank account on which the customer's cash flows arrive and the column *Probability of Default* gives the probability of default per year for the corresponding customer.

**Table 2.** The key data of the customers used in the case study.

<i>Customer</i>	C1	C2	C3	C4	C5	C6
<i>Currency</i>	EUR	EUR	USD	GBP	CHF	CHF
<i>Bank Account</i>	B1-1	B1-2	B1-3	B1-4	B2-1	B1-2
<i>Probability of Default</i>	0.5%	1.3%	1.1%	0.8%	0.9%	5.0%

<sup>1</sup> The exchange rates for September 10, 2016 ([www.oanda.com/currency/converter](http://www.oanda.com/currency/converter)).

We suppose that the currency of suppliers S1 and S2 is EUR, the currency of suppliers S3 and S4 is USD, and the currency used by suppliers S5 and S6 is GBP. These are the currencies in which the corresponding supplier expects the cash flows from the company to him. There is no fixed bank account set for each supplier and the treasurer of the company chooses the actual bank account [16].

**Cash Flows:** We generated, randomly, the incoming and outgoing cash flows.

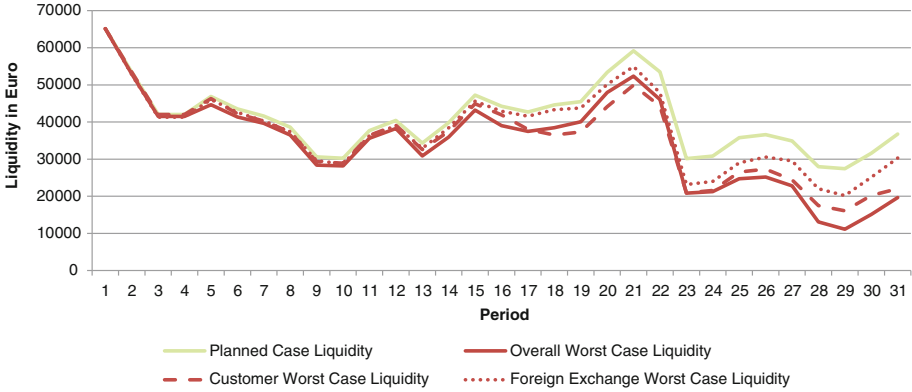
*Generation of Incoming Cash Flows:* We generated 60 *incoming* cash flows. We drew the value  $icf$  of each cash flow randomly from a normal distribution with  $\mathcal{N}(2400, 500^2)$ . For each incoming cash flow, we drew the origin  $r(icf)$  randomly from all customers given in Table 2. We set the bank account  $a$  and the currency  $\bar{c}(icf)$  as specified in Table 2. We drew a random due period  $t_{icf_a}^*$  as an integer between 1 and 29. We set the probability of default  $pd(icf_{a,t^*})$  to a uniformly distributed random number between 0% and 30%, rounded to two decimals. Further, we set the most likely delay  $M_{icf_{a,t^*}}$  randomly to either 0 or 1 periods and we set the maximum delay  $U_{icf_{a,t^*}}$  to a random integer between 3 and 5 periods. We set the minimal delay  $L_{icf_{a,t^*}}$ , as well as the minimal loss  $min_{icf_{a,t^*}}$ , to 0 for all incoming cash flows. Additionally, we set the probable loss  $mode_{icf_{a,t^*}}$  to a random value between 0 and  $(1/2) \cdot icf$ . Finally, we set the maximum loss  $max_{icf_{a,t^*}}$  with a 99% chance to a random value between  $(1/2) \cdot icf$  and  $icf$ . With a 1% chance, we set  $max_{icf_{a,t^*}}$  to  $icf$ .

*Generation of Outgoing Cash Flows:* We generated 60 *outgoing* cash flows. The value  $ocf$  was uniformly distributed between 300 and 5000. Both the supplier and the bank account  $a$  were drawn randomly and the currency was set according to the customer. Finally, we drew a random due period  $t_{ocf_a}^*$  between 1 and 29.

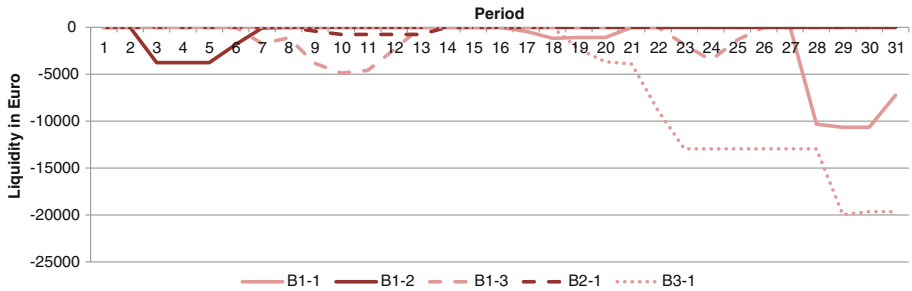
## 5.2 Results

**Risk Simulation:** Figure 1 depicts the planned development of the liquidity and the expected worst case development of the liquidity under risk over time. The continuous line at the top shows the planned liquidity development. The dotted line represents the development under foreign exchange risk, the dashed line indicates the liquidity development under the risk of customer default, and the continuous line at the bottom depicts the liquidity development taking the overall risk into consideration. The risk of bank defaults is omitted from Fig. 1 because in our case study, the risk turned out to be EUR 0. In order to calculate the risk of customer default, we simulated  $10000 \cdot 60 = 600000$  Monte-Carlo scenarios that took 154.6 s. In this example, the risk of customer default is the highest individual risk with a final value of EUR 14699.91 in the last period. We simulated  $4 \cdot 10000 = 40000$  scenarios to determine the foreign exchange risk, which took 20.5 s. The final value for the foreign exchange risk is EUR 6425.79 in the last period.

The estimation of the overall risk took  $(60 + 4 + 3) \cdot 10000 = 670000$  iterations and 180.3 s. The final value for the overall risk is EUR 17095.95. It is important



**Fig. 1.** The expected worst case liquidity induced by all three risk types simultaneously in comparison to the expected case liquidity.



**Fig. 2.** Checking the violation of constraints (2) by bank accounts.

to note that the overall risk is much lower than the sum of the three individual risks (EUR 14699.91 + EUR 6425.79 + EUR 0 = EUR 21125.70). This difference is justified by two facts: First, the probability that both the 95%-worst foreign exchange rates and the 95%-worst customer defaults happen in the same scenario is smaller than 5% and thus does not happen in the 95%-worst scenario. Second, there exist interdependencies between the foreign exchange risk and the other two risks since the foreign exchange rates have a high impact on all cash flows, both incoming and outgoing, in foreign currencies.

We simulated  $3 \cdot 10000 = 30000$  scenarios, required 12.4s, for computing the risk of bank default. Due to the short planning horizon of 30 days and the realistic probabilities of default, no bank goes bankrupt within the planning horizon. Thus, this risk remains EUR 0.

**The Constraints:** We are interested to know which bank accounts (and in which periods) violated constraints (2). For this purpose, Fig. 2 shows which bank account violates the constraints (2) as well as how much cash is missing in that specific period in order to satisfy constraints (2). As an example, consider

the bank account  $B1-2$ . In period 3, there are EUR 3761.31 missing to satisfy the constraints (2). In period six, the missing amount decreases to EUR 1829.91. The missing amount shrinks further to EUR 66.2 in period seven and diminishes completely in period eight. The bank account  $B1-4$  never violates the constraints (2) and is therefore omitted from Fig. 2.

## 6 Conclusion

In this paper, we propose a bi-objective optimization model that considers four distinct types of financial risk: the risk of default of customers, the risk of default of banks, the foreign exchange risk, and the overall risk. Furthermore, we propose a Monte-Carlo-based simulation algorithm to calculate the Cash Flow at Risk and give an illustrative example.

In order to use the presented model in a real-world context, we will need to do more experiments. Furthermore, it would be interesting to development an algorithm that automatically finds ways to improve the distribution of cash and cash flows among bank accounts.

## References

1. Avila-Pacheco, J.V., Morabito, R.: Application of network flow models for the cash management of an agribusiness company. *Compt. Ind. Eng.* **6**, 848–857 (2011)
2. Baumol, W.J.: The transactions demand for cash: an inventory theoretic approach. *Q. J. Econ.* **66**, 545–556 (1952)
3. Cooper, R.: *Corporate Treasury and Cash Management*. Palgrave Macmillan, Hampshire (2013)
4. Dentz, M., Arons, S.: Die Gunst der Stunde - Die Härte der Krise birgt die Chance der Bewährung. *Der Treasurer* **3**, 6–7 (2009)
5. Diamond, D.W., Dybvig, P.H.: Bank Runs, Deposit Insurance, and Liquidity. *J. Polit. Econ.* **91**, 401–419 (1983)
6. Golden, B., Liberatore, M., Lieberman, C.: Models and solution techniques for cash flow management. *Comput. Oper. Res.* **6**, 13–20 (1979)
7. Gormley, F.M., Meade, N.: The utility of cash flow forecasts in the management of corporate cash balances. *EJOR* **182**, 923–935 (2007)
8. Gregory, G.: Cash flow models: a review. *Omega* **4**, 643–656 (1976)
9. Hagerer, P.: *Cash Flow at Risk und Value at Risk in Unternehmen*. Universität Siegen, Cologne (2010)
10. Hinderer, K., Waldmann, K.-H.: Cash management in a randomly varying environment. *EJOR* **130**, 468–485 (2001)
11. Hormuth, M.W.: *Recht und Praxis des konzernweiten Cash-Managements: ein Beitrag zur Konzernfinanzierung*. Duncker & Humblot, Berlin (1998)
12. Hoseini, A., Andalib, R., Gatmiri, B.: Stochastic framework for cash flow forecasting considering owner's delay in payment by use of Monte Carlo simulation. In: 51st ASC Annual International Conference Proceedings (2002)
13. Miller, M.H., Orr, D.: A model of the demand for money by firms. *Q. J. Econ.* **80**, 413–435 (1966)

14. Schierenbeck, H., Lister, M., Kirmße, S.: Ertragsorientiertes Bankmanagement: Band 1: Messung von Rentabilität und Risiko im Bankgeschäft. Springer, Wiesbaden (2013)
15. Seethaler, P., Haß, S., Brunner, M.: Ermittlung und Aggregation von WährungsrisikenSeethaler. In: Seethaler, P., Steitz, M. (eds.) Praxishandbuch Treasury Management, pp. 343–362. Gabler, Frankfurt am Main (2007)
16. Treasury Intelligence Solutions. <https://www.tis.biz/>

# Demand Side Management: A Case for Disruptive Behaviour

Dina Subkhankulova<sup>1</sup>, Artem Baklanov<sup>2,3,4</sup>, and David McCollum<sup>2</sup>

<sup>1</sup> University College London Energy Institute,  
Central House, 14 Upper Woburn Place, London WC1H 0NN, UK  
`dina.subkhankulova.13@ucl.ac.uk`

<sup>2</sup> International Institute for Applied Systems Analysis,  
Schlossplatz 1, 2361 Laxenburg, Austria  
`{baklanov,mccollum}@iiasa.ac.at`

<sup>3</sup> N.N. Krasovskii Institute of Mathematics and Mechanics,  
Yekaterinburg, Russia

<sup>4</sup> Ural Federal University, Yekaterinburg, Russia

**Abstract.** The UK electricity system is undergoing a significant transformation. Increasing penetration of renewable generation and integration of new consumer technologies (e.g. electric vehicles) challenge the traditional way of balancing electricity in the grid, whereby supply matches demand. Demand-side management (DSM) has been shown to offer a promising solution to the above problem. However, models proposed in literature typically consider an isolated system whereby a single aggregator coordinates homogeneous consumers. As a result potential externalities of DSM are overlooked. This work explores the value of DSM in the context of an interacting electricity system, where utilities compete for cheap electricity in the wholesale market. A stylized model of the UK electricity system is proposed, whereby a traditional supplier competes with a ‘green’ supplier in the wholesale market. The modelling was able to show that with enough dispatchable capacity the traditional supplier was able to benefit from instructing his consumers to increase demand peaks, which had an adverse effect on the system.

**Keywords:** Demand side management · Competing utilities · UK electricity system

## 1 Introduction

Climate policy amongst other triggers such as lowering costs for ICT, storage and micro generation technology are driving changes within the UK power system.

On the supply side, the UK has seen a significant growth in the deployment of renewable power generators over the last decade, in particular wind and solar [12]. On the demand side, a number of technologies have been entering the market, such as small scale batteries, electric vehicles [4], heat pumps and micro-generation units (especially rooftop solar PV) [13]. Consumers are also becoming more active due to increasing proliferation of smart power metering and

management technology. The government is planning to equip every household with smart meters for electricity and gas by 2020 [3]. The changes on the supply and demand sides of the electricity system are causing concern for the grid, as it becomes more difficult to coordinate variable supply with unpredictable demand.

Demand side management (DSM) can offer a promising solution to balancing the electricity grid. Certain technologies like electric vehicles or electrical storage can be scheduled to operate during times of high renewable supply. A number of coordination methodologies have been proposed in the literature (see Sect. 2). However, such models typically ignore the interactions between the aggregators in the wholesale market. In reality, electricity suppliers compete in the wholesale market for cheap electricity. It is then possible to imagine that electricity companies may manipulate consumer demand in order to gain a competitive advantage. Consequently, it becomes uncertain how this will impact the system as a whole.

This project is concerned with investigating how DSM may impact the security of the grid. We pose the following question: “*Are there conditions under which DSM can be disruptive to the grid?*”. In order to answer it we develop a stylised model for the UK electricity grid, whereby two types of suppliers (a traditional and green) compete in the wholesale electricity market.

In the following report Sect. 2 gives an overview of the previous work in the domain of DSM and provides motivation for the project; Sect. 3 describes the proposed model; Sect. 4 covers model calibration and initial set-up; Sect. 5 provides result interpretation and analysis; and Sect. 6 gives a summary of the conclusions and suggestions for areas of improvement.

## 2 Relevant Work

The idea of using demand side flexibility to compensate for intermittent supply is not new [14]. However, due to the lack of communication technology the work remained preliminary and thus untested in simulation settings. Recent developments in communication and data management tools (smart meters, mobile internet, cloud computing) alongside rapid integration of renewables have reignited academic interest in demand-side control as a means to compensate for variable supply. The new DSM models assume the presence of software agents which can optimise electricity usage on behalf of the consumer. Compared to the traditional DSM schemes aimed at human behaviour, software agents are able to perform complex calculations faster using tools such as machine learning and optimisation. There is a large body of research focusing on different ways of performing DSM (see [1, 18]). A major shortcoming of these models is that the system under consideration often represents an idealistic setting where a set of homogeneous consumers are being coordinated by a single aggregator, e.g. [7, 17]. On the other hand whole system models like in [15] tend to assume perfect consumer and market behaviour in order to perform global optimisation. Consequently, the dynamic interactions between autonomous consumers and suppliers are lost.

In reality, electricity suppliers interact in the wholesale market in order to supply consumers with very different demand profiles and flexibility resources.

Following these gaps in research, we propose a dynamic model which would highlight the benefits and potential issues concerning DSM in the context of the wholesale electricity market.

### 3 The Model

The following model is motivated by the recent changes in the retail electricity market. Two types of electricity suppliers compete for cheap electricity: a vertically integrated utility owning dispatchable power generation capacity (TS)<sup>1</sup> and a ‘new’ independent supplier owning renewable generation capacity. We will refer to the later as the ‘green supplier’ (GS)<sup>2</sup>. At this stage there exist two agents representing each type: TS and GS. The consumers are modelled to possess small scale batteries with Tesla Power Wall specifications due to commercial availability of the technology [16].

Whereas GS can offer greener electricity, it is unable to fulfil its consumer demand without going to the market (where TS profits from selling electricity). On the other hand, TS can choose to reserve its capacity for supplying its own consumers instead of selling it in the market (in which it may lose out on profit opportunity in the market). Finally, both suppliers may utilise smart coordination mechanisms in order to influence the flexible demand of their consumers (see Sect. 3.1). The two suppliers compete on the retail price they can offer to the consumers which is calculated as the break-even cost of supplying them with electricity<sup>3</sup>.

#### 3.1 The Agents

The following model runs on 24h basis with all decisions being made a day ahead. We define a daily period counter  $i = 1, \dots, I$ , where  $I$  stands for the total number of daily periods (here 24h) and a day counter  $t = 1, \dots, T$ , where  $T$  stands for the total number of simulated days (here 365 or one year). The following section describes the intraday calculations where we drop the day counter  $t$  for clarity of notation.

**Consumers.** Consumer agents represent residential households. We consider a set of consumers  $\mathcal{A}$ , where each agent  $a \in \mathcal{A}$  has a daily demand profile,  $d_i^a(t)$ <sup>4</sup>. The demand profile can be split into a non-deferrable part ( $b_i^a$ ) and a flexible part ( $f_i^a$ ), s.t.

$$d_i^a = b_i^a + f_i^a, \quad \forall i \in [1, I],$$

where  $f_i^a$  represents the battery storage profile and is calculated as the difference between charging and discharging profiles  $f_i^{a+}$  and  $f_i^{a-}$ . Conceptually,

<sup>1</sup> This type of a suppliers represents one of the ‘Big Six’ energy utilities operating in the UK market.

<sup>2</sup> These companies represent the new entrants in the UK electricity market like Ecotricity and Good Energy.

<sup>3</sup> This serves as a step for further development of the model where the consumers are able to switch suppliers.

<sup>4</sup> We use standard residential demand profiles provided by the National Grid [5].



non-deferrable demand corresponds to those activities where consumption cannot be shifted in time such as cooking or watching TV. Flexible demand can be shifted in time subject to consumer storage specifications, which in our case include battery storage capacity  $e^a$  (kWh), minimum and maximum charging power constraints,  $f_{min}^a$  and  $f_{max}^a$  (kW), and efficiency  $\eta^a$ .

Each day the supplier may deploy a coordination strategy, in which case it will negotiate the storage profile with the consumer prior to physical electricity delivery (See Sect. 3.1).

**Suppliers.** Suppliers are energy companies responsible for providing their consumers with electricity. We use index  $S \in \{T, G\}$  to differentiate between the traditional (TS) and the green supplier (GS). Hence, we identify two subsets of consumers:  $\mathcal{A}^T \subseteq \mathcal{A}$  (those signed up with TS) and  $\mathcal{A}^G \subseteq \mathcal{A}$  (those signed up with GS). We assume that the size of the two consumer sets is the same, i.e.  $|\mathcal{A}^T| = |\mathcal{A}^G|$ .

Suppliers' objective is to fulfill energy demand of its consumers,  $B_i^S$ , which is calculated as the sum of individual consumer demand profiles, i.e.

$$B_i^S = \sum_{a \in \mathcal{A}^S} d_i^a \quad \forall i \in [1, I] \quad \text{and} \quad S \in \{T, G\}.$$

The suppliers may do so by generating electricity from their own resources of capacity,  $cap^S$ , and/or by buying electricity in the market. If the supplier generates  $R_i^S$  electricity in a daily period  $i$ , the net demand left to be fulfilled from the market becomes

$$D_i^S = \max(0, B_i^S - R_i^S) \quad \forall i \in [1, I] \quad \text{and} \quad S \in \{T, G\}.$$

Now, an assumption is made that GS does not sell electricity in the market since it wants to maximise the use of its own renewable resources. Thus, for GS,  $R_i^G$  is constrained purely by the installed capacity of the wind generator,  $cap^G$ , and the weather<sup>5</sup>. The traditional supplier has an option of selling electricity in the market in which case it will not use all of self-generated electricity. Hence, for TS,  $R_i^T$  is constrained by the self-utilisation parameter,  $u(t)$ , and installed generation capacity  $cap^T$ , i.e.

$$R_i^T = \min(cap^T * u(t), B_i^T) \quad \forall i \in [1, I],$$

where we make a rational decision assumption that the TS will not use more than it needs to fulfill consumer demand.

Consequently, the amount of electricity left for the TS to sell in the market becomes

$$Q_i^T = cap^T - R_i^T.$$

The traditional supplier sells electricity in the wholesale market at a price  $z$  calculated as

$$z^T = p_{SRMC}^T + \epsilon^T, \tag{1a}$$

<sup>5</sup> In order to model the generation capacity of the GS we take historical electricity supply profile from a 1.8MW wind farm in Wales [9, 10].

where  $\epsilon^T$  is referred to as an ‘uplift’ and represents any additional costs incurred by the supplier excluding the cost of running the generator, e.g. transmission and distribution costs, and company operation.

The first term in (1a),  $p_{SRMC}^T$ , is known as the short run marginal cost (SRMC) of generator type  $T$  and is calculated as

$$p_{SRMC}^T = c_{var}^T + \frac{p_{fuel}^T}{\eta^T}, \quad (1b)$$

where,

$c_{var}^T$  = variable operating and maintenance cost for a generator of supplier  $T$ ,

$p_{fuel}^T$  = price of fuel used by supplier  $T$  to generate electricity,

$\eta^T$  = efficiency of an electricity generator of supplier  $T$ .

For the green supplier (1b) is reduced to  $p_{SRMC}^G = c_{var}^G$  since wind is free.

Finally, at the end of day  $t$ , both types of suppliers calculate the break-even retail price of electricity:

$$\pi^S(t) = \frac{\sum_{i=1}^I ((R_i^S(t) + Q_i^{sold,S}(t))p_{SRMC}^S(t) - Q_i^{sold,S}(t)z^S(t) + D_i^S(t)p_i(t))}{B_i^S(t)}, \quad (2)$$

where,

$R_i^S(t)$  = electricity generated for self-use by supplier  $S$  in daily period  $i$  of day  $t$ ,

$Q_i^{sold,S}(t)$  = electricity sold by supplier  $S$  in daily period  $i$  of day  $t$ ,

$z_i^S(t)$  = the asking price for a unit of energy by a supplier  $S$  in day  $t$ ,

$p_i(t)$  = the market price for a unit of energy in daily period  $i$  and day  $t$ , and

$B_i^S(t)$  = the total power supplied to the consumers by a supplier  $S$  in day  $t$ .

We can split (2) into three parts: cost of running the generator, profit made in the market and the cost of purchasing additional electricity<sup>6</sup>. Calculating the retail price at break-even cost enables us to compare the competitiveness of the two utilities. Since we only have one TS in the model, index  $S$  is dropped from  $z^S$  and  $u^S$  for the rest of the report.

**TS Learning.** As will be seen in the next section, it is critical for TS to set the offer and self-utilisation parameters. Hence, we propose Algorithm 1 to allow the TS to learn the best strategy. The algorithm is based on the method developed by [2] which uses reinforcement learning to teach the agents the best strategy to adopt in the market in terms selecting the offer price ( $z(t)$ ) and the self-utilisation parameter ( $u(t)$ ). The idea is that the agent experiments with strategies for the first ten simulation days, after which the exploration time is reduced to 50% with the remaining time dedicated to selecting the best available strategy.

**Supplier Coordination.** We consider two decentralised coordination strategies, whereby the supplier is signaling its consumers on how to schedule storage. The two algorithms are based on the method developed by [7] but have

<sup>6</sup> Since the GS does not sell electricity in the market it omits the second term from (2).

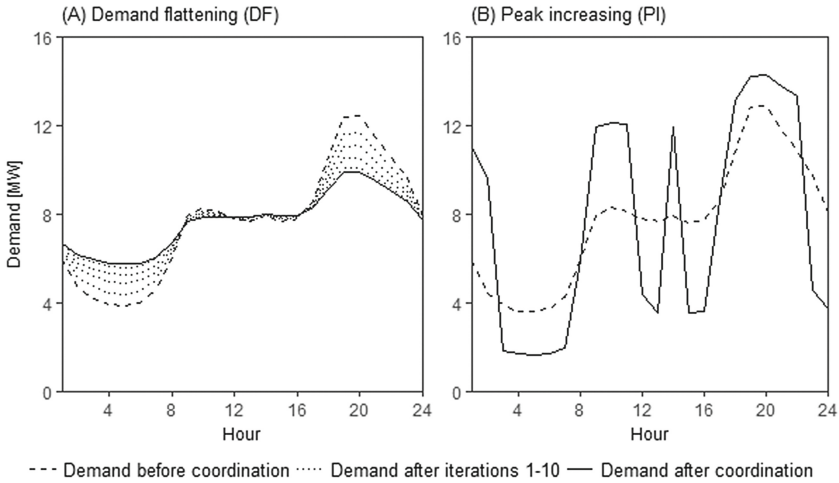
**Algorithm 1.** TS learning algorithm

**Require:** Retail price  $\pi(t)$ , offer  $z(t)$  and self-utilisation parameter  $u(t)$  from day  $t$  and the matrix for storing results,

$$\mathcal{M} = \begin{bmatrix} \pi(1) & u(1) & z(1) \\ \vdots & \vdots & \vdots \\ \pi(10) & u(10) & z(10) \end{bmatrix}.$$

**Ensure:** Supplier self-utilisation parameter  $u(t+1)$  and offer  $z(t+1)$ .

- 1: Generate five random values  $a_1, a_2, a_3, a_4, a_5$ , s.t.  
 $a_1, a_2, a_3, a_4 \in \mathcal{U}(0.9, 1.1)$  and  $a_5 \in \mathcal{U}(0, 1)$
  - 2: **if**  $t \leq 10$  **then**
  - 3:    $\mathcal{M}_{t,1} \leftarrow \pi(t), \mathcal{M}_{t,2} \leftarrow u(t), \mathcal{M}_{t,3} \leftarrow z(t)$
  - 4:    $u(t+1) \leftarrow u(t) * a_1, z(t+1) \leftarrow z(t) * a_2$
  - 5: **else**
  - 6:   Sort the strategy matrix  $\mathcal{M}$  in order of ascending retail prices, s.t.  
 $\forall m_{k1}, m_{(k+1)1} \in \mathcal{M}, \quad m_{k1} \leq m_{(k+1)1} \forall, \quad k \in [1, 10]$
  - 7:   **if**  $a_5 \leq 0.5$  **then**
  - 8:      $u(t+1) \leftarrow \mathcal{M}_{1,2}, z(t+1) \leftarrow \mathcal{M}_{1,3}$
  - 9:   **else**
  - 10:    $u(t+1) \leftarrow u(t) * a_3, z(t+1) \leftarrow z(t) * a_4$
- return**  $u(t+1), z(t+1)$



**Fig. 1.** Demonstration of the two decentralised coordination mechanisms deployed in the model by suppliers.

been adapted to consumers with batteries rather than electric vehicles. The first coordination algorithm is designed to reduce the variance of the demand profile and hence we call it ‘demand flattening’ or DF (Fig. 1, left). The second coordination algorithm is designed to increase the variance of the demand profile and hence we call it ‘peak increasing’ or PI (Fig. 1, right). The supplier negotiates the demand profile with the consumers over a number of iterations whilst the consumers imposes electrical storage constraints.

### 3.2 The Market

The market represents a pool of electricity generation companies which sell power. These may be independent generators or vertically integrated companies also possessing a retail business (like in the case of TS). The generators bid available capacity into the market at a set offer per unit of energy. The cheaper units of electricity get sold first with more expensive units reserved for times of higher electricity demand. Hence, electricity prices are positively correlated with system's demand for electricity. The market is cleared at the price of the marginal unit of electricity – the last unit of generation needed to fulfill system demand.

In this model, the market receives the sum of the electricity demand profile from the two suppliers in each time period  $i$  and day  $t$ , i.e.  $L_i(t) = D_i(t)^T + D_i^G(t)$ . The market price,  $p_i(t)$ , is set at a level of the last unit of marginal generation capacity needed to fulfil the demand,  $L_i(t)$ . For simplicity, we assume a linear relationship between system demand and prices:

$$p_i(t) = k \times L_i(t). \quad (3)$$

The TS is able to sell electricity in the market, however only if the price offered,  $z(t)$ , is lower than the clearing price,  $p_i(t)$  as calculated in (3). If the offer is too high, the supplier is unable to sell and uses electricity for self-consumption. On the other hand, if the supplier bids too low, it misses out on a profit opportunity. In order to decide on the best strategy the TS deploys the learning Algorithm 1 as described in Sect. 3.1.

## 4 Experimental Set-Up

The model described in this paper is highly stylised and hence, in order to capture real-life interactions, it was critical to set the parameters.

Firstly, it was decided that 30,000 consumers (aggregated to 30 modelled agents) was a sufficient number to capture the needed model interactions without compromising on the speed of the simulation. Consumers were equally split between the TS and GS in order to make the retail prices comparable. The regression coefficient in (3) was adjusted to 4.2 in order to achieve the same level of market prices compared to the historical value of £40/MWh in the *base case* – when no utility coordinated (meaning that storage was not operated) [6].

The SRMC for a traditional supplier was set at £14/MWh (to represent a coal power plant) and at £1.5/MWh for the green supplier. It was assumed that 50% of consumers signed with each supplier were in possession of a battery with capacity of 6.4kWh, charging power of 3kW and 100% efficiency. It was also necessary to cap the demand peak value in the PI algorithm to 1.2 of the max daily residual demand value before coordination, as a value above that led to the TS losing out by paying more in the market.

The purpose of this project was two-fold. Firstly, to investigate under which strategies the suppliers benefited through offering a lower retail price compared to their opponent. Secondly, whether any of the strategies resulted in a negative effect to the system through increased demand peaks. Hence, we consider six

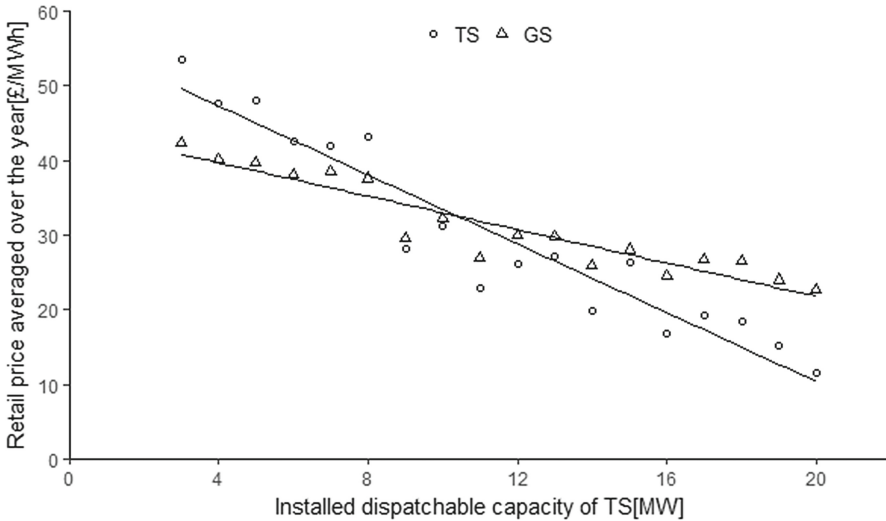
combinations of supplier coordination strategies referred to as ‘cases’ (Table 1). We adopt the notation whereby the first item in the brackets represents the strategy adopted by the TS (in capital letters) and the second item represents GS strategy (in small letters).

**Table 1.** Matrix representation of the simulation cases.

GS strategy	TS strategy		
	<i>NC</i>	<i>DF</i>	<i>PI</i>
<i>NC</i>	<i>(NC, nc)</i>	<i>(DF, nc)</i>	<i>(PI, nc)</i>
<i>DF</i>	<i>(NC, df)</i>	<i>(DF, df)</i>	<i>(PI, nc)</i>

During the preliminary run it was found that the model was very sensitive to the amount of installed generation available to suppliers. Figure 2 demonstrates how the retail prices for both suppliers change when the TS acquires more dispatchable capacity. In order to keep the experiment fair we swooped through a range of parameters for installed capacities for both suppliers during the experiment – each corresponding to a scenario.

For each run we compare the competitiveness of the two suppliers by tracking their average retail prices for the year,  $\pi^T(t)$  and  $\pi^G(t)$ . In order to assess the impact on the system we also monitor the system demand,  $L_i(t)$  – a proxy for carbon intensity of the grid and an indicator for the security of the electricity transmission system. Please refer to Appendix A for the overall model flow<sup>7</sup>.



**Fig. 2.** Comparison of retail prices for TS and GS under different scenarios of installed dispatchable capacity for TS, Case: (PI, nc).

<sup>7</sup> The model code is also available on request.

## 5 Results and Analysis

Suppliers benefited from having more generation capacity. With 10 MW, TS achieved a lower retail price compared to GS five cases out of six (Fig. 3, left). Out of the five cases the highest retail price was obtained under PI coordination (PI, nc) corresponding to a more profitable strategy. Increasing demand peaks enabled the TS to sell electricity in the market above SRMC and increase the overall price level whilst still keeping the retail price competitive. For GS,

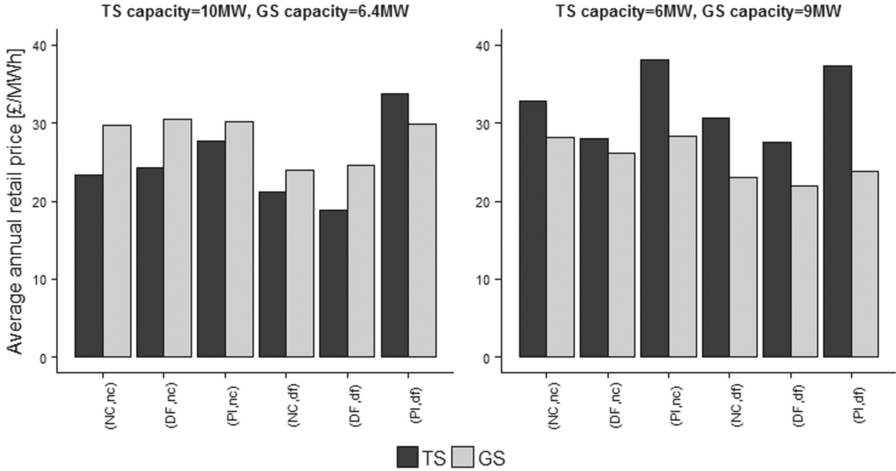


Fig. 3. Comparison of average annual retail prices achieved by suppliers.

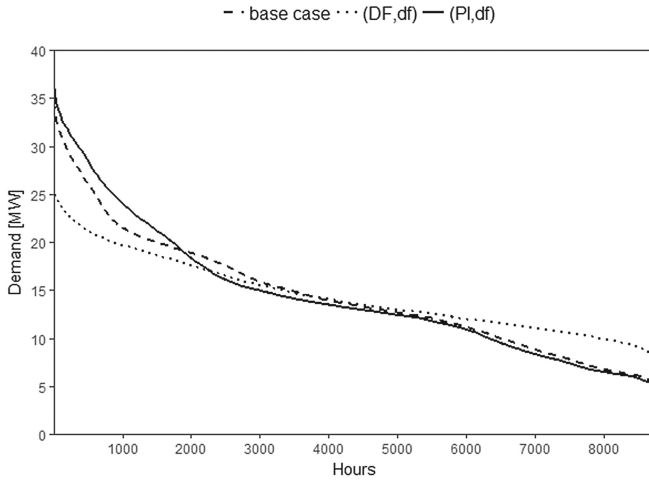


Fig. 4. Comparison of duration curves for three experimental cases.

increased renewable capacity meant more independence from the market and hence lower prices for consumers (Fig. 3, right).

In terms of system performance, suppliers benefited from deploying demand flattening coordination. This led to lower system demand and consequently electricity prices for both supplier (Fig. 4, Case (DF, df)).

In fact peak demand reached 26.7 MW compared to the 34.7 MW for the base case scenario when no coordination was performed (Fig. 4, Case (NC, nc)). Peak increasing strategy carried a negative effect of increased system demand leading to a peak demand of 36.6 MW (Fig. 4, Case (PI, df)).

## 6 Conclusion and Further Work

In this work we investigated how a traditional supplier (TS) and a green supplier (GS) can compete for consumers by utilising DSM strategies. We considered a number of scenarios by varying the parameters of supplier generation capacity. In each scenario we investigated the outcomes under different combination of supplier strategies by monitoring retail prices and system demand.

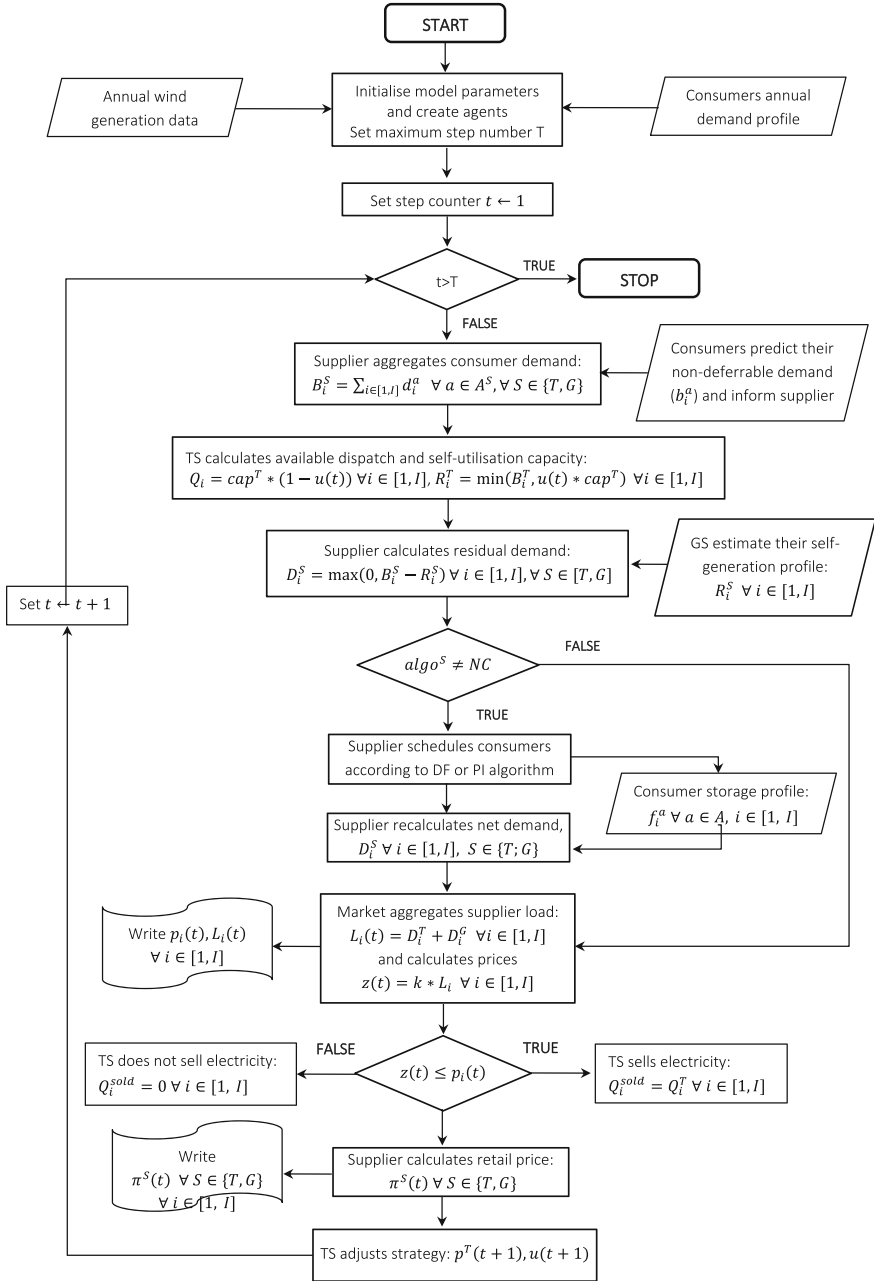
The modelling was able to show that with enough dispatchable capacity, conditions existed when the TS benefited from instructing its consumers to increase demand peaks by offering a lower retail price as compared to GS. PI coordination led to a higher retail price and hence more profitable conditions for TS (assuming that operational costs did not change). This suggests a possibility that such a strategy could be considered by a utility. Nevertheless, PI strategy resulted in an adverse effect of increased demand peaks in the system. In order to compete, the green supplier was obliged to perform demand flattening coordination. There are a number of limitations which we aim to address in the future:

- Increase the number of suppliers bidding in the market
- Equip consumers with the ability to choose supplier
- Introduce better learning algorithm to the suppliers
- Introduce uncertainty to the demand and supply sides
- Introduce heterogeneous consumers
- Equip consumers with the ability to generate electricity

DSM can offer a promising solution in balancing variable supply with flexible demand and help transition the UK electricity system to a cleaner more sustainable one. However, if not controlled it could lead to negative effects for the whole system. Thus, in order to extract the maximum amount of benefit out of DSM a relevant regulatory framework is likely to be required in the future. For example, electricity supplier pricing strategies could be regularly reviewed so as to ensure that they are not incentivizing consumer behavior that works against overall electricity system performance.

**Acknowledgments.** Part of the research was developed in the Young Scientists Summer Program at the International Institute for Systems Analysis, Laxenburg (Austria) with financial support from the United Kingdoms National Member Organization.

## A Model Flow Diagram





## References

1. Bomann, T., Eser, E.J.: Model-based assessment of demand-response measures. A comprehensive literature review. *Renew. Sustain. Energy Rev.* **57**, 1637–1656 (2016). <http://doi.org/10.1016/j.rser.2015.12.031>
2. Bower, J., Bunn, D.W.: Model-based comparisons of pool and bilateral markets for electricity. *Energy J.* **21**(3), 1–29 (2000). <https://www.jstor.org/stable/41322889>
3. DECC. Smart meters: a guide. London (2016). <https://www.gov.uk/guidance/smart-meters-how-they-work>
4. DfT Vehicle Licensing Statistics. Electric vehicle market statistics (2016). <http://www.nextgreencar.com/electric-cars/statistics/>
5. Demand profiling-ELEXON (2013). <https://www.elexon.co.uk/reference/technical-operations/profiling/>
6. Elexon. Market Index Data - BMRS (2015). <https://www.bmreports.com/bmrs/?q=balancing/marketindex/historic>
7. Gan, L., Topcu, U., Low, S.H.: Optimal decentralized protocol for electric vehicle charging. *IEEE Trans. Power Syst.* **28**(2), 940–951 (2013). <http://doi.org/10.1109/TPWRS.2012.2210288>
8. Gan, L., Wierman, A., Topcu, U., Chen, N., Low, S.H.: Real-time deferrable load control: handling the uncertainties of renewable generation. In: Fourth International Conference on Future Energy Systems (E-Energy 2013), pp. 113–124 (2013). <http://doi.org/10.1145/2567529.2567553>
9. Pfenninger, S., Staffell, I.: Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. *Energy* **114**, 1251–1265 (2016). doi:10.1016/j.energy.2016.08.060. <https://www.renewables.ninja>
10. Staffell, I., Pfenninger, S.: Using bias-corrected reanalysis to simulate current and future wind power output. *Energy* **114**, 1224–1239 (2016). doi:10.1016/j.energy.2016.08.068. <https://www.renewables.ninja>
11. Electricity supply market shares by company: Domestic (GB). Ofgem (2016). <https://www.ofgem.gov.uk/chart/electricity-supply-market-shares-company-domestic-gb>
12. Renewable UK. UK Wind Energy Database (UKWED) (2016). <http://www.renewableuk.com/en/renewable-energy/wind-energy/uk-wind-energy-database/index.com>
13. Renewable Energy Foundation (REF). Energy Data (2016). <http://www.ref.org.uk/energy-data>
14. Schweppe, F., Daryanian, B., Tabors, R.D.: Algorithms for a spot price responding residential load controller. *IEEE Trans. Power Syst.* **4**(2), 507–516 (1989). <http://www2.econ.iastate.edu/tesfatsi/ResidentialLoadControllerAlgorithms.FSchweppeEtAl1989.pdf>
15. Strbac, G., Aunedi, M., Pudjianto, D., Djapic, P., Gammons, S., Druce, R.: Understanding the Balancing Challenge (2012). [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/48553/5767-understanding-the-balancing-challenge.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/48553/5767-understanding-the-balancing-challenge.pdf)
16. Tesla. Powerwall - The Tesla Home Battery (2017). <https://www.tesla.com/en-GB/powerwall>

17. Voice, T., Vytelingum, P., Ramchurn, S., Rogers, A., Jennings, N.: Decentralised control of micro-storage in the smart grid. In: AAAI 2011: Twenty-Fifth Conference on Artificial Intelligence, pp. 1421–1426 (2011). <http://eprints.soton.ac.uk/272262/>
18. Yang, Z., Li, K., Foley, A., Zhang, C.: Optimal scheduling methods to integrate plug-in electric vehicles with the power system: a review. In: IFAC World Congress, pp. 8594–8603 (2014)

# Enhancing Reduced Order Model Predictive Control for Autonomous Underwater Vehicle

Prashant Bhopale<sup>(✉)</sup>, Pratik Bajaria, Navdeep Singh, and Faruk Kazi

Department of Electrical Engineering, VJTI, Mumbai 400019, India  
bhopleprashant@outlook.com, pratikbajaria@live.com,  
{nmsingh,fskazi}@vjti.org.in

**Abstract.** Performing control actions to complete the mission depending on the feedback measurements is a crucial task in case of Autonomous Underwater Vehicle due to dependency on navigation system. For such task a Reduced Order Model Predictive Control (ROMPC) has been implemented using highly nonlinear model of AUV to control motion in the dimensional space under the assumption that the feedback measurements at every iteration are clearly available to solve the quadratic problem. But in real-time scenario, navigation system collects measurements from the sensors installed on the hardware part of the AUV which may fail due to vulnerability of sensors to onboard equipment noise or poor signal during diving operation resulting in failure of ROMPC furthermore mission, hence proper state estimator or observer is required for real time operation to support navigation system. This work proposes a solution, based on the optimal estimation property of Extended Kalman Filter in the presence of process and measurement noise or missing measurements to estimate position and orientation of AUV for successful and enhanced feedback control application.

**Keywords:** Autonomous Underwater Vehicle · Reduced Order Model Predictive Control · Extended Kalman Filter · Navigation system · Sensor noise

## 1 Fixed-Period Problems: The Sublinear Case

Oceans are the central engine of energy and chemical balance that sustains mankind. They provide warmth and power; they deposit hydrocarbon and mineral resources that are important to mankind need; they moderate the weather so food can be grown on the land to feed the earth's population; their living resources also supply food directly [1]. All of these necessitate the mankind to investigate, analyze and protect the oceans and to develop means to explore the oceans. For this purpose underwater vehicles are widely used.

Underwater vehicles are categorized in manned or unnamed underwater vehicle. Autonomous underwater vehicles (AUVs) are unmanned type underwater vehicle which are more mobile, have much wider reachable scope, and can achieve

real-time control. While their capacity is limited by onboard power sources, memory devices, intelligence and lots of relative technologies are still in lab research stage. Although AUVs have promising applications with the development of scientific and technological means. AUVs are capable of traveling underwater without requiring input from an operator and equipped with energy sources such as batteries or fuel cells, different navigation sensors like inertial measurement unit (IMU), sonar and pressure sensor, GPS and so on, and embedded controllers which are pre-programmed for missions. Hence various sensors developed on AUV provide the position and velocity measurements of AUV which includes Doppler based sensor for linear velocity hence position (in surge, sway and Yaw direction) also IMUs (Internal Measurement Units) used for angular position and acceleration (in Roll, Pitch and Yaw direction). Although magnetic compass and GPS can be used to provide absolute position but underwater reception of such signals is weak hence AUV has to surfaced many times to correct the GPS position during the mission. Alternatively to deal with this problem Long-Baseline (LBL) Acoustic Positioning System is used to correct GPS position using calculation of relative position of the two vehicles using but it increases the complexity of the navigation system. Similarly position of the vehicle can also be computed using Simultaneous Localization and Mapping (SLAM) techniques [2] with respect to a map of the environment. Since the direct measurement of the vehicle's position is difficult and SLAM methods are not always applicable, hence underwater navigation depends strongly on method called dead reckoning. Given the AUV's initial position and orientation, the current positions and orientation of AUV is estimated by integrating measurements from the IMUs or similar navigation sensors and the Doppler Velocity Log (DVL) [3].

However, in the case of a sensor failure or sensor measurement corruption by the noise generated from onboard equipment like rudder or stern motors, magnetic interference results in 2 effects. First the AUV has no means to navigate further, as a result, the success of the AUV's mission is compromised, besides, without accurate navigation or position, it may be impossible to retrieve the AUV. Second different feedback control algorithms like PID, LQR, MPC shown in [6–9], Robust  $H_\infty$  Control as [10], Hybrid Control as [11] - and so on have been implemented on AUV which are obvious state or measurement dependent for close loop control, hence noise corrupted or missing measurements may result in close loop control failure, infeasible solution or control action finally resulting in mission failure or loss of AUV. To deal with such noise or disturbances Constrained or Tube base MPC control have been implemented [12–14] but it increases the computation cost and have limitation as disturbances bounded. Hence a controller-observer scheme is required to deal with such noise and bounded control contains with comparatively lower computation cost.

As a remedy the nonlinear state estimator or a observer can be used to estimate the states in case of corrupted measurements or predict the states in case of missing measurements [4] which enhances the control algorithm for nonlinear

model. Hence this paper proposes a Enhancing Model Predictive Control of AUV using Extended Kalman Filter for 6 DoF (Degree of Freedom) and 12 order non-linear model.

The paper structured as follow: Sect. 1 states the nonlinear mathematical model of AUV with a brief introduction with Kinematics and kinetics. Section 2 states and explains the proposed enhancing scheme for ROMPC control of AUV by defining reduced order model and model predictive control, and application of EKF in presence of noisy or missing measurements. Section 3 shows justification of proposed method with help of simulation of 6 DoF nonlinear AUV model control enhanced using EKF by estimating feedback states for ROMPC. Conclusion and future work is presented in Sect. 4.

## 2 Mathematical Modelling of Autonomous Underwater Vehicle

A mathematical model of a generic Autonomous Underwater Vehicle (AUV) with 6 Degree of Freedom (6-DoF) via 12 first order equations is referred in this paper. According to SNAME [5] the notation for marine vessel, where the linear position and Euler angles (Angular Positions) components as shown in Fig. 1 refereed form [15] where surge, sway, heave, roll, pitch and yaw are  $[x, y, z, \phi, \theta, \psi]$  and corresponding linear and angular velocity are expressed as  $[u, v, w, p, q, r]$  receptively.

The nonlinear kinematics and kinetics can be derived by defining body fixed reference frame  $[X_B, Y_B, Z_B]$  and the earth fixed reference frame  $[X_E, Y_E, Z_E]$  as shown in [4].

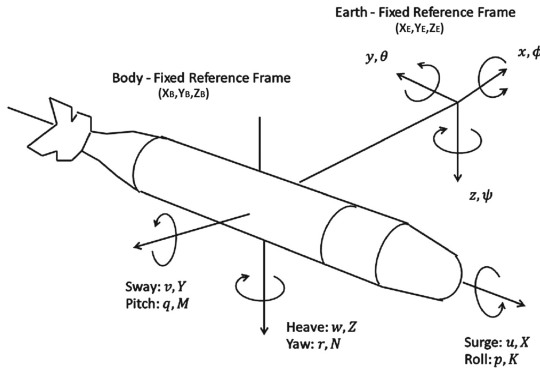


Fig. 1. Motion in 6 DoF considering earth fame and body frame

### 2.1 Kinematics

Kinematics deals with geometric aspects of motion of body, in our case AUV. The 6 Dof standard kinematic equations of motion can be written in component form as [4].

$$\begin{aligned}
 \dot{x} &= u\cos(\psi)\cos(\theta) + v[\cos(\psi)\sin(\theta)\sin(\phi) - \sin(\psi)\cos\phi] \\
 &\quad + w[\sin(\psi)\sin(\phi) + \cos(\psi)\cos(\phi)\sin(\theta)] \\
 \dot{y} &= u\sin(\psi)\cos\theta + v[\cos(\psi)\cos(\phi) + \sin(\phi)\sin(\theta)\sin(\psi)] \\
 &\quad + w[\sin(\theta)\sin(\psi)\cos(\phi) - \cos(\psi)\sin(\phi)] \\
 \dot{z} &= -u\sin(\theta) + v\cos(\theta)\sin(\phi) + w\cos(\theta)\cos(\phi) \\
 \dot{\phi} &= p + q\sin(\phi)\tan(\theta) + r\cos(\phi)\tan(\theta) \\
 \dot{\theta} &= q\cos(\phi) - r\sin(\phi) \\
 \dot{\psi} &= q\frac{\sin(\phi)}{\cos(\theta)} + r\frac{\cos(\phi)}{\cos(\theta)}, \theta \neq \pm 90^\circ
 \end{aligned} \tag{1}$$

## 2.2 Kinetics

The rigid-body kinetics can be derived by applying Newtonian mechanics in as following, first three equations represent the translational motion, while the last three equations represent the rotational motion.

Standard Kinetic equations of motion for submerged vehicle shown below,

Surge, or translation along the x-axis is given by:

$$m [\dot{u} - vr + wq - x_g(q^2 + r^2) + y_g(pq - \dot{r}) + z_g(pr + \dot{q})] = \sum X_{ext} \tag{2}$$

Sway, or translation along the y-axis is given by:

$$m [\dot{v} - wp + ur - y_g(r^2 + p^2) + z_g(qr + \dot{p}) + x_g(qp + \dot{r})] = \sum Y_{ext} \tag{3}$$

Heave, or translation along the z-axis is given by:

$$m [\dot{w} - uq + vp - z_g(p^2 + q^2) + x_g(rp - \dot{q}) + y_g(rq + \dot{p})] = \sum Z_{ext} \tag{4}$$

Roll, or rotation about the x-axis is given by:

$$I_{xx}\dot{p} + (I_{zz} - I_{yy})qr + m[y_g(\dot{w} - uq + vp) - z_g(\dot{v} - wp + ur)] = \sum K_{ext} \tag{5}$$

Pitch, or rotation about the y-axis is given by:

$$I_{yy}\dot{q} + (I_{xx} - I_{zz})rp + m[z_g(\dot{u} - vr + wq) - x_g(\dot{w} - uq + vp)] = \sum M_{ext} \tag{6}$$

Yaw, or rotation about the z-axis is given by:

$$I_{zz}\dot{r} + (I_{yy} - I_{xx})pq + m[x_g(\dot{v} - wp + ur) - y_g(\dot{u} - vr + wq)] = \sum N_{ext} \tag{7}$$

Where  $\sum x_{ext}, \sum Y_{ext}, \sum Z_{ext}, \sum K_{ext}, \sum M_{ext}$  and  $\sum N_{ext}$  are external forces added in Surge, Yaw, Sway, Roll, Pitch and Yaw directions due to the drag, added mass, lift, etc. forces. For the limitation of the space the nomenclature and derivation is not explained in this paper, interested readers can refer [4] for the same. The entire nonlinear 12 order and 6 DoF equation of motion combining Eqs. (1)–(7) can be represented in vector as

$$\begin{aligned}
 \eta_{k+1} &= f(\eta_k, \delta_k) \\
 \xi_k &= h(\eta_k)
 \end{aligned} \tag{8}$$

where  $\eta = [x, y, z, \phi, \theta, \psi, u, v, w, p, q, r]$  and  $\delta = [\delta_s, \delta_r]$ .

### 3 Proposed Method

The proposed methodology uses two independent reduced order models for horizontal and vertical control derived from nonlinear model in Eq. 8 by neglecting weakly coupled dynamics as [15] and shown in Sect. 2.1. Then MPC algorithm is used for designing the controller in horizontal and vertical plane, the control input generated  $(\delta_s, \delta_r)$  i.e. rudder and stern angle respectively are used to control the nonlinear plant of AUV. The enhancing of the reduced order MPC control stated above can be explained using block diagram shown in Fig. 2 below, where the MPC controller gets access to the actual plant states using EKF as a observer or state estimator by removing process and measurement noise there by reducing the chances of AUV mission failure due to process and measurement noise or missing measurements.

#### 3.1 Reduced Order Models

For simplicity we assume constant surge rate, and the complete model from Eqs. (1)–(7) is reduced into two controllable subsystems as in [15], these non interacting systems are given below:

**Horizontal Control (Surge, Yaw and Yaw Rate).** The states considered to obtain the reduced dynamics in horizontal plane are  $\eta_H = [v, r, \psi]^T$  assuming  $u = 1.54$  m/s and other states as zero and the reduced model thus obtained is shown in (11).

$$\begin{aligned}
 A_H(v, r, \psi) &= \left. \frac{\partial f}{\partial \delta_r} \right|_{u=1.54; w, p, q, \psi, \theta=0} \\
 &= \begin{bmatrix} -131|v| - 131v \text{sign}(v) - 58.89 & 37.51 + 0.6r + 0.6|r| & 0 \\ -3.18|v| - 3.18v \text{sign}(v) + 16.35 & -9.4r \text{sign}(r) - 9.4|r| - 6.05 & 0 \\ 0 & 1 & 0 \end{bmatrix} \\
 B_H(v, r, \psi) &= \left. \frac{\partial f}{\partial \delta_r} \right|_{u=1.54; w, p, q, \psi, \theta=0} \\
 &= [22.86 \quad -14.585 \quad 0]^T
 \end{aligned} \tag{9}$$

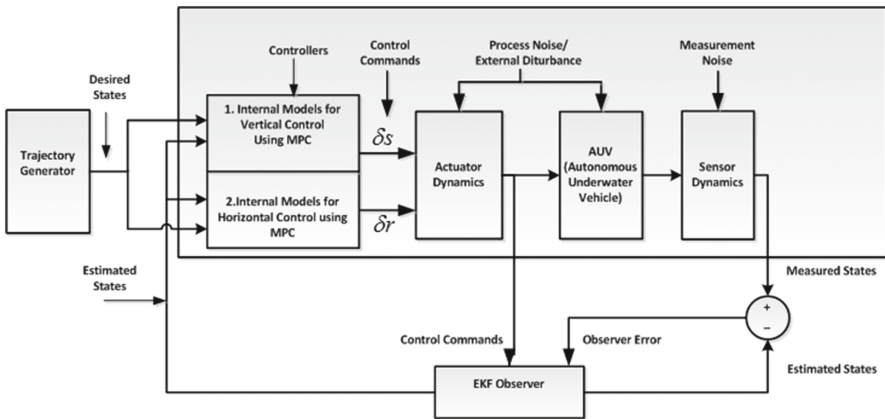


Fig. 2. Proposed method

**Vertical Control (Depth, Depth Rate, Pitch and Pitch Rate).** The reduced dynamics in horizontal plane is derived by considering states  $\eta_V = [w, q, \theta, z]^T$  and the obtained reduced model is shown in (12).

$$\begin{aligned}
 A_V(w, q, \theta, z) &= \left. \frac{\partial f}{\partial \eta_V} \right|_{u=1.54; v, p, r, \phi, \psi=0} \\
 &= \begin{bmatrix} -131w \operatorname{sign}(w) & 37.51 + 1.2q & 0 & 0 \\ -131|w| - 58.89 & -0.6q \operatorname{sign}(q) - 0.6|q| & & \\ -0.6q + 3.18w \operatorname{sign}(w) & -0.6w - 9.4q \operatorname{sign}(q) & -5.87c\theta + 182.87s\theta & 0 \\ +3.18|w| - 16.35 & -9.4|q| - 6.05 & & \\ 0 & 1 & 0 & 0 \\ c\theta & 0 & -1.54c\theta - ws\theta & 0 \end{bmatrix} \\
 B_V(w, q, \theta, z) &= \left. \frac{\partial f}{\partial \delta_r} \right|_{u=1.54; v, p, r, \phi, \psi=0} \\
 &= [-22.86 \ -14.585 \ 0 \ 0]^T
 \end{aligned} \tag{10}$$

The discretized dynamics with sampling time  $T_s$  at current states  $\eta(k)$  are given as,

$$\eta_i(k+1) = A_{i_d} \eta_i(k) + B_{i_d} \delta_j(k) \tag{11}$$

where  $A_{i_d} = (I + A_{i_d} T_s)|_{\eta_i=\eta_i(k)}$  and  $B_{i_d} = (B_{i_d} T_s)|_{\eta_i=\eta_i(k)}$ ,  $\eta_i(k) \in \mathfrak{R}^n$ ,  $\delta_j(k) \in \mathfrak{R}^m$  for  $i \in \{H, V\}$  and  $j \in \{r, s\}$ .

### 3.2 MPC for Reduced Order Models

MPC is based on iterative, finite-horizon optimization of a plant model as shown in Fig. 3 refereed form [15].

We consider the discretized dynamics with sampling time  $T_s$  at current state  $\eta(k)$  as,

$$\eta_i(k+1) = A_{i_d} \eta_i(k) + B_{i_d} \delta_j(k) \tag{12}$$

with control input constraint as

$$\delta_j(\min) \leq \delta_j \leq \delta_j(\max). \tag{13}$$

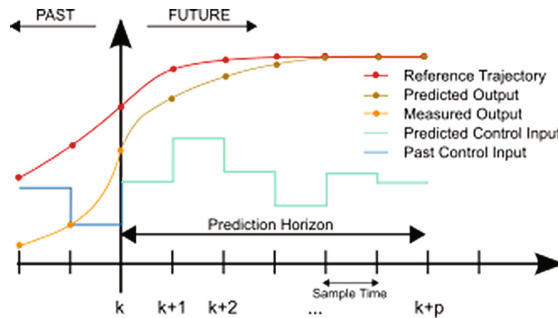


Fig. 3. Basic MPC scheme.



The cost function considered for finite horizon optimal control problem subjected to (12) is given by

$$\begin{aligned} \min J(\eta_{i_0}, u) &= (\eta_{i_N} - \eta_{i_{ref_N}})^T \bar{P} (\eta_{i_N} - \eta_{i_{ref_N}}) \\ &+ \sum_{k=0}^{N-1} (\eta_{i_k} - \eta_{i_{ref_k}})^T \bar{Q} (\eta_{i_k} - \eta_{i_{ref_k}}) + \delta_{jk}^T \bar{R} \delta_{jk} \end{aligned} \quad (14)$$

where  $\bar{P}$ ,  $\bar{R}$  are positive definite matrix and  $\bar{Q}$  is positive semi-definite matrix.  $\eta_i$  is  $i^{th}$  state of state vector  $\eta \in \mathbb{R}^n$  and  $\delta_i$  is  $i^{th}$  state of state vector  $u \in \mathbb{R}^m$ .

Optimal problem stated above is solved using batch approach. According to this approach the optimal control problem is converted into a multi-parametric programming problem with vector parameter.

For control horizon  $N$  we can write

$$\underbrace{\begin{bmatrix} \eta_{i_1} \\ \eta_{i_2} \\ \vdots \\ \eta_{i_N} \end{bmatrix}}_{\eta} = \underbrace{\begin{bmatrix} A_{id} \\ A_{id}^2 \\ \vdots \\ A_{id}^N \end{bmatrix}}_S x_{i_0} + \underbrace{\begin{bmatrix} B_{id} & 0 & \dots & 0 \\ A_{id} B_{id} & B_{id} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{id}^{N-1} & A_{id}^{N-2} & \dots & B_{id} \end{bmatrix}}_T \underbrace{\begin{bmatrix} u_{j_0} \\ u_{j_1} \\ \vdots \\ u_{j_{N-1}} \end{bmatrix}}_{\delta} \quad (15)$$

Equation 15 can be rewritten as

$$\eta = S\eta_{i_0} + T\delta \quad (16)$$

Thus the cost function (14) can be written as

$$J(\eta_{i_0}, \delta) = (\eta - \eta_{ref})^T \hat{Q} (\eta - \eta_{ref}) + \delta^T \hat{R} \delta \quad (17)$$

where

$$\hat{Q}_{Nn \times Nn} = \text{blockdiag}[\bar{Q}, \bar{Q}, \dots, \bar{Q}],$$

$$\hat{R}_{Nm \times Nm} = \text{blockdiag}[\bar{R}, \bar{R}, \dots, \bar{R}] \text{ and}$$

$\eta_{ref}$  is a vector of reference trajectories over horizon. Here  $n$  is number of states and  $m$  is a number of control inputs.

Substituting (16) into (17) we get

$$J(\eta_0, \delta) = \delta^T H \delta + F \delta \quad (18)$$

where  $H = S^T \hat{Q} S + \hat{R}$  and  $F = 2(\eta_0^T T^T - \eta_{ref}) \hat{Q} S$ .

The cost function (18) becomes the quadratic function of  $\delta$  (Control input). This optimization problem is solved using the quadratic programming approach to find the control input  $\delta$ .

### 3.3 Extended Kalman Filter (EKF)

For linear Gaussian systems, it can be shown that the closed-form of the optimal recursive solution is given by the well known Kalman Filter. It is widely used

in sensor and navigation systems since it can reconstruct unmeasured states as well as remove white and colored noise from the state estimates. It is also possible to include wild-point removal capabilities. In case of soft temporarily loss of measurements, the filter equations behave such as a predictor. As soon as new measurements are available, the predictor is corrected and updated online to give the minimum variance estimate. This feature is particularly useful when satellite signals are lost since the Kalman filter can predict the motion using only gyros and accelerometers.

For real world applications, which are neither linear nor Gaussian, the Bayesian approach is intractable. Thus, a sub-optimal approximated solution has to be found. For instance, in the case of a nonlinear system with non-Gaussian probability density function (pdf), the Extended Kalman Filter (EKF) algorithm approximates a nonlinear system with its truncated Taylor series expansion around the current estimate. Here, the nonGaussian pdf is approximated with its first two moments, mean and covariance. The key assumption when designing a EKF here is that the system model is observable, this is necessary in order to obtain convergence of the estimated states  $\hat{\eta}$  to  $\eta$ . Moreover, if the system model is observable, the state vector  $\eta \in \mathfrak{R}^n$  can be reconstructed recursively through the measurement vector  $\xi \in \mathfrak{R}^m$  and the control input vector  $\delta = [\delta_r, \delta_s]^T \in \mathfrak{R}^c$  as discussed in [4]. Proposed Enhancing algorithm is shown in Fig. 2 to incorporate EKF.

The discrete-time Extended Kalman filter (Kalman, 1960) is defined in terms of the discretized system model for Eq. 8 by considering process noise  $\omega_k$  and measurement noise  $\epsilon_k$  with convenience  $P_k$  and  $Q_k$  at instant  $k$ ; and can be written as:

$$\begin{aligned}\eta_{k+1} &= f(\eta_k, \delta_k) + \omega_k \\ \xi_k &= h(\eta_k) + \epsilon_k\end{aligned}\tag{19}$$

where expected mean values of process and measurement noise are  $E[\omega_k] = 0$  and  $E[\epsilon_k] = 0$  respectively.

The model parametric recursive estimation algorithm for EKF is stated below,

### Prediction Step

States and convenience matrix are predicted at  $k + 1^{th}$  instant by using estimated vales at  $k^{th}$  instant.

$$\begin{aligned}\hat{\eta}_{k+1} &= f(\eta_k, \delta_k) \\ \hat{P}_{k+1} &= F_k P_k F_k^T + Q_k\end{aligned}\tag{20}$$

### Correction Step

Kalman gain is calculated as,

$$K_k = \hat{P}_{k+1} H^T \left[ H \hat{P}_{k+1} H^T + R_k \right]^{-1}\tag{21}$$

Estimated states and process noise convenience matrix are corrected as,

$$\begin{aligned}\eta_{k+1} &= \hat{\eta}_{k+1} + K_k (\xi_k - H\hat{\eta}_{k+1}) \\ P_{k+1} &= \hat{P}_{k+1} - K_k H \hat{P}_{k+1}\end{aligned}\quad (22)$$

where the state transition and observation matrices are defined to be the following Jacobians

$$\begin{aligned}F_k &= \left. \frac{\partial f(\eta, \delta)}{\partial \eta} \right|_{\hat{\eta}_k, \delta_k} \\ H_k &= \left. \frac{\partial h(\eta)}{\partial \delta} \right|_{\hat{\eta}_k}\end{aligned}\quad (23)$$

The use of EKF can be seen in variety applications, in [16] it has been shown that the continuous-time EKF is incremental GES under the assumption that the  $P_i$  matrix of the Riccati equation is uniformly positive definite and upper bounded, that is

$$p_{min}I \leq P_i \leq p_{max} \quad (24)$$

for two strictly positive constants  $p_{min}$  and  $p_{max}$ . This guarantees that the estimates converge exponentially to the actual states.

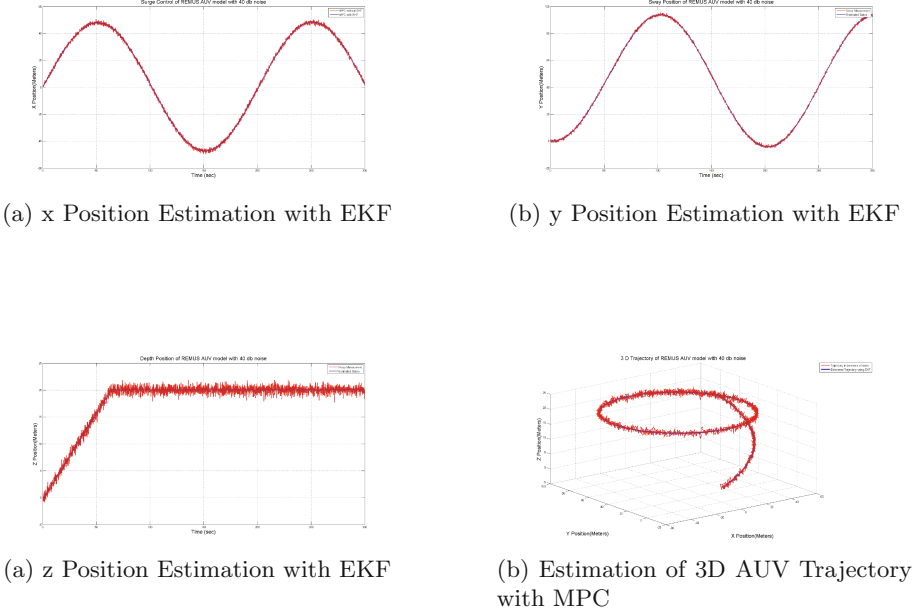
**Assumption -A1** [*Bypass corrector in case of Dead-reckoning*]. *During sensor failures, the best thing to do is to trust the model without any updates. Hence, the corrector is bypassed by setting  $\eta_{k+1} = \hat{\eta}_{k+1}$  in Eq. (22) and prediction is based on the system model only:*

$$\eta_{k+1} = f(\eta_k, \delta_k) \quad (25)$$

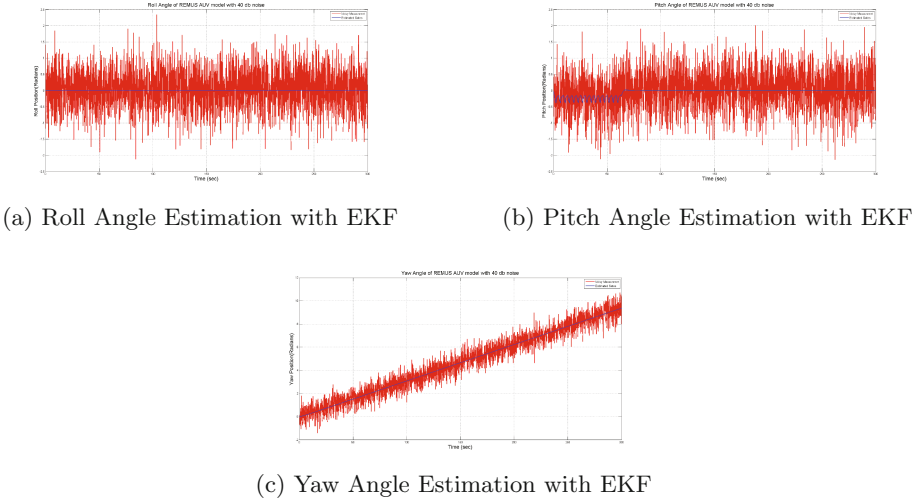
**Remark -A1.** *Dead-reckoning refers to the case where there are no updates (signal loss) for a period of time. In such scenario  $\xi_k$  will be unavailable in Eq. (22) which will keep increasing the convenience and make prediction go wrong way. Hence in such case the assumption -A1 removes the kalman gain factor and make controller rely on the system model's prediction based on last available measurement and correcting the estimation as soon as measurements gets available. This assumption solve the problem arises due to missing or lost of measurement getting from IMU or GPS and enhance the system performance.*

## 4 Simulation Results

Simulation of the proposed Enhanced Reduced order MPC strategy for the nonlinear AUV model is carried out in MATLAB and analysis along performance in presence of noise for proposed scheme in discussed in this section. For the simulation purpose we used the nonlinear AUV model along with system parameter are taken from [17]. The underactuated system considered for the AUV is assumed to have independent surge speed control which maintain the forward velocity of 1.54 knots/s.



**Fig. 4.** Estimation of linear position in presence of noise



**Fig. 5.** Estimation of angular position in presence of noise

For the two independent control in horizontal and vertical plane we consider following design parameters: Control Horizon  $N = 5, \bar{R} = 0.1, \bar{Q} = \text{diag}[001]'$  for horizontal control and  $\bar{Q} = \text{diag}[0001]$  for vertical control. Positive definite matrix  $\bar{P}$  is obtained by solving discrete algebraic Riccati equation for final

cost as shown in Eq. (17). The constraints on control input of rudder and stern angle are assumed as  $-30 \leq \delta_r \leq 30$  and  $-20 \leq \delta_s \leq 20$ , respectively. The initial covariance matrices for EKF are assumed as  $P_0 = \text{diag}[\text{ones}(12, 1)]$  and  $Q_0 = \text{diag}[0.1 * \text{ones}(12, 1)]$ . Depth reference of 20 m and yaw angle reference of ramp with 1.8 deg/s slope are considered. Also 40 db measurement noise has been added as shown in Fig. 2 and Eq. (19) to see the performance of EKF. With the help of designed parameters and simulation references, the proposed scheme is simulated to control nonlinear model of AUV in three-dimensional space in presence of noise and outputs of simulations results are illustrated in Figs. 4 and 5.

## 5 Conclusion

In this paper, enhancing algorithm for reduced order MPC in presence of sensor noise or missing measurements for underactuated nonlinear model of AUV is proposed to control AUV in 3 dimensional space. The proposed algorithm used two reduced order model predictive controller and the feedback states given for the controller are filtered for additive noise added due to onboard equipment. The proposed scheme also reduces the chances of AUV mission failure due to lost measurements due in case of poor signal of GPS during diving action, failure of IMU or sensor failure to control AUV by trusting system model and successfully recover the AUV or complete the mission in such scenario.

The hardware in loop simulation and Graphical User Interface (GUI) designed for same have been developed and not included in this paper due to page limitation and real time application can be performed for the proposed scheme which formulates the future work or extension of this paper.

**Acknowledgments.** The authors would like to acknowledge the support of Centre of Excellence (CoE) in Complex and Nonlinear dynamical system (CNDS), through TEQIP-II, VJTI, Mumbai, India.

## References

1. National Research Council: Underwater Vehicles and National Needs. Academy Press, Washington, D. C. (1996)
2. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics. MIT Press, Cambridge (2005)
3. Filaretov, F., Zhirabok, A.N., Zyev, A.V., Protsenko, A.A., Tuphanov, I.E., Scherbatyuk, A.F.: Design investigation of dead reckoning system with accommodation to sensors errors for autonomous underwater vehicle, OCEANS 2015 - MTS/IEEE Washington, Washington, DC, pp. 1–4 (2015)
4. Fossen, T.I.: Handbook of Marine Craft Hydrodynamics and Motion Control. Wiley, Hoboken (2011)
5. Nomenclature for Treating the Motion of a Submerged Body Through a Fluid Technical and Research Bulletin No. 1–5. New York, Society of Naval Architects and Marine Engineers (SNAME)

6. Lee, J., Roh, M., Lee, J., Lee, D.: Clonal selection algorithms for 6-DOF PID control of autonomous underwater vehicles. doi:[10.1007/978-3-540-73922-716](https://doi.org/10.1007/978-3-540-73922-716)
7. Khodayari, M.H., Balochian, S.: Modeling and control of autonomous underwater vehicle (AUV) in heading and depth attitude via self-adaptive fuzzy PID controller. *J. Mar. Sci. Technol.* **20**, 559 (2015). doi:[10.1007/s00773-015-0312-7](https://doi.org/10.1007/s00773-015-0312-7)
8. Joo, M.G., Qu, Z.: An autonomous underwater vehicle as an underwater glider and its depth control. *Int. J. Control Autom. Syst.* **13**, 1212 (2015). doi:[10.1007/s12555-014-0252-8](https://doi.org/10.1007/s12555-014-0252-8)
9. Pereira, F.L., de Sousa, J.B., Gomes, R., Calado, P., van Schuppen, J.H., Villa, T.: A model predictive control approach to AUVs motion coordination. In: *Coordination Control of Distributed Systems*. Springer International Publishing (2015). doi:[10.1007/978-3-319-10407-2\\_2](https://doi.org/10.1007/978-3-319-10407-2_2), 978-3-319-10407-2
10. Nag, A., Patel, S.S., Kishore, K., Akbar, S.A.: A robust  $H_\infty$  based depth control of an autonomous underwater vehicle. In: *2013 International Conference on Advanced Electronic Systems (ICAES)*, Pilani, pp. 68–73 (2013)
11. Marco, D.B., Healey, A.J., McGhee, R.B.: Autonomous underwater vehicles: hybrid control of mission and motion. doi:[10.1007/978-1-4613-1419-6-5](https://doi.org/10.1007/978-1-4613-1419-6-5)
12. Bumroongsri, P.: Tube-based robust MPC for linear time-varying systems with bounded disturbances. *Int. J. Control Autom. Syst.* **13**, 620 (2015). doi:[10.1007/s12555-014-0182-5](https://doi.org/10.1007/s12555-014-0182-5)
13. Kouvaritakis, B., Cannon, M.: *Control, Non-linear Predictive: Theory and Practice*, vol. 61. Iet (2001)
14. Kouvaritakis, B., Cannon, M.: *Model Predictive Control: Classical, Robust and Stochastic*. Springer, Heidelberg (2015)
15. Jagtap, P., Raut, P., Kumar, P., Gupta, A., Singh, N., Kazi, F.: Control of autonomous underwater vehicle using reduced order model predictive control in three dimensional space”. *IFAC-PapersOnLine* **49**(1), 772–777 (2016)
16. Jouffroy, J., Fossen, T.I.: A tutorial on incremental stability analysis using contraction theory. *Model. Identif. Control MIC* **31**(3), 93–106 (2010)
17. Prestero, T.: Development of a six-degree of freedom simulation model for the REMUS autonomous underwater vehicle. In: *MTS/IEEE Conference and Exhibition in OCEANS*, vol. 1, pp. 450–455 (2001)

# Comparison of Feedback Strategies for Supporting Programming Learning in Integrated Development Environments (IDEs)

Jarno Coenen<sup>(✉)</sup>, Sebastian Gross, and Niels Pinkwart

Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany  
{jarno.coenen, sebastian.gross,  
niels.pinkwart}@hu-berlin.de

**Abstract.** In this paper we investigate whether providing feedback to learners within an Integrated Development Environment (IDE) helps them write correct programs. Here, we use two approaches: feedback based on stack trace analysis, and feedback based on structural comparisons of a learner program and appropriate sample programs. In order to investigate both approaches, we developed two prototypical extensions for the Eclipse IDE. In a laboratory study, we empirically evaluated the impact of the extensions on learners' performance while they solved programming tasks. The statistical analyses did not reveal any statistically significant effects of the prototype extensions on the performance of the learners, however, the results of a qualitative analysis imply that the provided feedback had at least a marginal impact on the performance of some learners. Also, feedback from the participants confirmed the benefit of providing feedback directly within IDEs.

**Keywords:** Adaptive feedback · Integrated Development Environment · Java programming

## 1 Introduction

PROGRAMMING skills are important in both professional and educational contexts. Computer science education below the university level is encouraged by a number of governments in Europe [1, 2], as well as in the USA [3]. University level enrolment in computer science studies is also steadily increasing each year [4]. Learning a programming language can, among other means, be effectively aided by solving smaller programming tasks. This learning method can be assisted by teachers (tutors) and/or tools. Here, the application of computer-supported learning systems can be effective in time and financially sensitive situations. There are several applications of computer-supported learning systems. For instance, Intelligent Tutoring Systems (ITSs) simulate the role of human tutors using models of formalized domain knowledge and student knowledge to provide feedback to learners when solving a given programming problem [5].

Learners often struggle with identifying and fixing errors in their programs. When a stack trace is printed in the console, indicating one or more errors, it is often difficult for

novice programmers to understand [6–9]. Integrated Development Environments (IDEs) are a common approach for support, not only for experienced programmers, but also for learners to find and fix errors in their programs by implementing a wide range of features such as syntax highlighting, syntax checks, debugging, and minor semantic and logical checks. However, even though IDEs provide different features to address syntactical and semantical issues, not all of them are appropriate for novices. For instance, debugging a program requires a programmer to have knowledge about the mistake(s) she made. We therefore propose to extend IDEs by implementing features which address both logical and semantic mistakes, as well as syntactical errors in learners' programs, and thus help them develop correct programs.

The outline of this paper is as follows: In Sect. 2, we review related work in the domain of learning programming. We introduce our approach and summarize its design and implementation in Sect. 3. We conducted a lab study where we evaluated the proposed approach. A description of this study and its results are summarized in Sect. 4. Section serves as a conclusion and an outlook on future work.

## 2 Related Work

There are many approaches which aim to support learners of programming in using computer-supported learning systems. Ample research articles exist which focus on how to support novice learners in understanding stack traces. HelpMeOut [10] is a social recommender system for the Java-based IDE called Processing<sup>1</sup>. This system offers assistance to learners for one specific stack trace at a time, and recommends a bug fix based on a database of bug fixes from other peers. DrScheme [7] is an IDE for the Lisp-based dialect Scheme. Using this system, a learner can select a language level (e.g., “Beginner Student Language”) and receive support messages for each of these errors, which serve as reminders for concepts that the learner is expected to know depending on the selected level. Gauntlet [11], on the other hand, is a pre-processing program that is only accessible online through a specific website. Learners can enter their Java programs through a web form after which they can view Gauntlet's outcome. In this outcome, possible syntax errors are then explained in pedagogically suitable phrasing, and some novice level semantic errors are also identified and sufficiently explained. From the given examples, it seems that Gauntlet simply alters compiler messages rather than offer additional error description text with examples.

Aside from stack trace based approaches, there are other systems that aim to help the learner learn programming using AI and adaptive techniques. J-LATTE [12] is an Intelligent Tutoring System (ITS) that provides two modes for teaching a subset of Java. The first mode is purely conceptual, involving the abstract algorithmic solving of problems without writing Java code. The second mode is a coding mode that introduces the use of Java code for the first time within the program. This mode provides on-demand feedback for constraint violation in the learner's program. This feedback is

---

<sup>1</sup> Processing is both a Java based IDE, as well as a kind of simplified Java programming language for educational purposes. The project website is available at <http://www.processing.org/>.



designed as text-based note for each violated constraint. JITS [13] is another ITS where a learner can extend fragments of a given program for a given exercise on her own. Here, a learner can request feedback on demand, for which the system provides syntactical help with quick fixes. To some extent, feedback for logical errors based on the exercise specification is also provided. JO-Tutor [14] is an ITS program which focus on how to handle Java object by providing digital learning materials, such as quizzes and sample erroneous programs, as well as a Java editor. The learner then receives feedback if their solution is correct.

All of these approaches have one crucial drawback: the user learns a programming language in an environment that is typically different from real world usage scenarios. When a learner moves to a real IDE, she is required to adapt her techniques accordingly. If the program simply alters errors for a learner, she is then also required to relearn the exceptions of the programming language in question. If instead a real IDE is used, with error messages compliant to the programming language standard, no transition to the real world usage is needed. Therefore, it is worth developing features that are seamlessly integrated into IDEs in order to provide a variety of types of feedback to learners. These features would be particularly helpful in providing feedback for solving semantic and logical errors, as even these basic programming skills are often insufficiently understood by learners [6–9]. This problem stems from the fact that erroneous but syntactically correct programs can be compiled and partly executed, but not necessarily produce the desired outcome. Here, feedback on semantic and logical errors could provide meaningful information to assist the learner in overcoming this initial hurdle.

### 3 Approach: Design and Implementation

As stated, our goal is to extend IDEs to provide meaningful feedback to learners in order to help them develop syntactically and semantically correct programs. We chose Eclipse IDE as the host system above other alternative IDEs (e.g. NetBeans<sup>2</sup>, and IntelliJ IDEA<sup>3</sup>), because it is more widely used among programmers [15], [16]<sup>4</sup>, [17, 18]. Due to its popularity in professional software development and computer science education at university level, we focused on supporting Java programmers.

In order to help learners find and fix both syntactical and logical/semantical mistakes, we developed two extensions for the Eclipse IDE: (i) in order to complete the programming cycle, a stack trace analysis provides feedback to a learner, and (ii) based on comparison of learner’s solution and appropriate sample solutions, we highlight and contrast both the learner’s solution and (steps of) a sample solution. The proposed design tries to be integrated seamlessly into the IDE, and guide the learner on demand through the available feedback.

---

<sup>2</sup> NetBeans is an open source IDE supporting, amongst others, Java (see <https://netbeans.org/>).

<sup>3</sup> IntelliJ IDEA is an Apache 2 licensed IDE for non-commercial use supporting, amongst others, Java (see <https://www.jetbrains.com/idea/>).

<sup>4</sup> In this survey in 2016, Eclipse is ranked second to IntelliJ for the first time in their survey, however only by a marginal 5%.

### 3.1 Feedback Based on Stack Trace Analysis

The first extension (referred to as EXT\_1) is based on analyzing stack trace exceptions from console output. When an exception occurs while executing a program, the extension parses the stack trace to retrieve the exception signature and the associated lines of code. The signature is then matched against the exceptions previously defined in a database. If the exception has been defined in the database, then the type of exception is known. In this case, three types of exceptions are possible: a runtime exception, a syntax exception, or a logic error. Using the information from the database, the extension then highlights related lines of the code in a window (as illustrated in Fig. 1). Each type of exception has its own distinctive background color, and unknown exceptions share one default background color. The code line throwing the exception is highlighted in a bright color, and subsequent associated code lines in a lighter shade. Highlighted lines also provide additional tooltips with notes from the database, and a button to request examples. Clicking on this button produces a pop-up dialog consecutively listing all examples defined in the database for this specific exception.

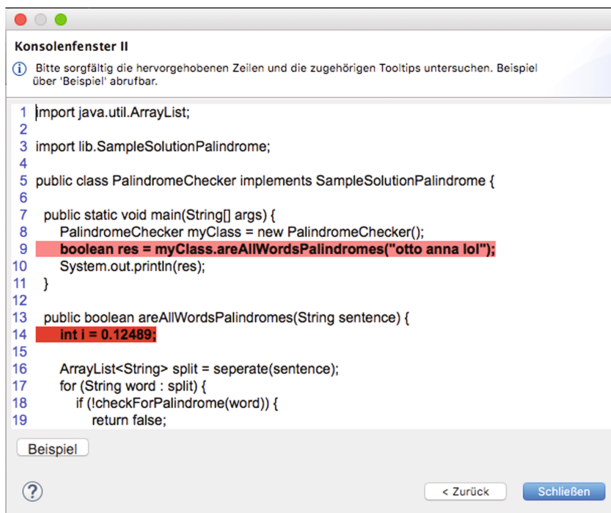


Fig. 1. View on stack trace analysis details.

For providing feedback to learners based on stack trace analysis, we composed a database of typical errors in Java. Since a variety of entries were possible, as Java offers a vast amount of possible errors and exceptions, we reviewed several papers to identify those errors which are typically made by novice programmers (see [19–25]). Our research produced five papers (referenced in Table 1) that included enough raw data to compose the following table, which consolidates each paper’s top 10 most commonly made errors. Overall, the literature research revealed 9 errors (see Table 1), which we added to the

stack trace database. We also added the additional errors ‘NullPointerException’, ‘ArrayIndexOutOfBoundsException’, ‘cannot cast from ... to ...’, and ‘cannot make a static reference to the non-static’.

**Table 1.** Java errors typically made by learners.

Error message	Papers
Cannot resolve symbol	[19, 20, 22–24]
; expected	[19, 20, 22–24]
Illegal start of expression	[19, 20, 22–24]
Bracket-expected (i.e. ‘(,’, ‘{,’, ‘[‘or’]’)	[19] <sup>a</sup> , [20, 22–24]
Class or interface expected	[19, 20, 22]
<identifier> expected	[19, 22, 24]
Incompatible types	[19, 20, 22]
Unknown-method	[20, 22, 23]
Else without if	[22, 24]
Not a statement	[19, 24]
Reached EOF while parsing	[23, 24]
Unknown-class	[20, 23]
.class expected	[20]
Missing return statement	[22]
Method cannot be applied to parameter types	[23]

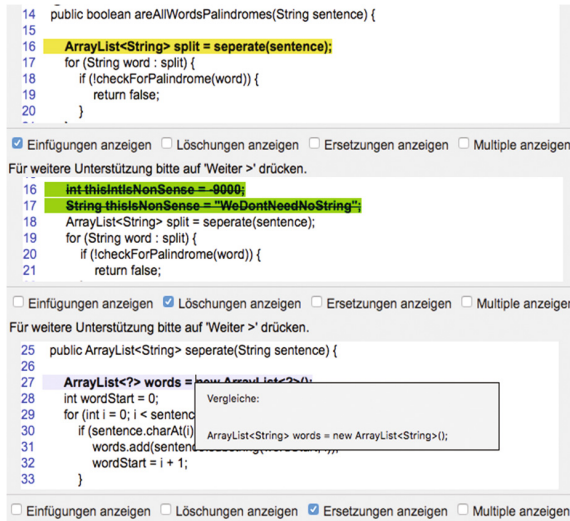
<sup>a</sup>These papers offer more detail, i.e., they list the errors separately, but are listed here only once for better comparability.

### 3.2 Feedback Based on Comparisons to Sample Programs

The second extension (referred to as EXT\_2) implements the example-based learning approach and uses feedback strategies whose benefits were evaluated in previous studies [26, 27]. Technically, this approach is based on the “TCS” toolbox<sup>5</sup> which analyzes Java programs and calculates a dissimilarity score for two Java programs [28]. The toolbox uses advanced machine learning techniques. Although the toolbox identifies the most similar sample solution, the extension can opt to choose the next sequential (more advanced) sample solution step for alignment and eventual presentation. If a more advanced sample solution step exists that is similar to both the originally identified sample solution step and to the learner solution, but not to the final sample solution, then this more advanced sample solution is used for alignment. This procedure is meant facilitate the learner in making further progress.

We implemented a view (illustrated in Fig. 2) which displays an instruction message prompting a learner to carefully examine highlighted code areas, and a text area containing the learner’s current Java editor code (see Fig. 2). Below the text area, there

<sup>5</sup> For more details, see the homepage of the toolbox, available here: <https://openresearch.cit-ec.de/projects/tcs>.

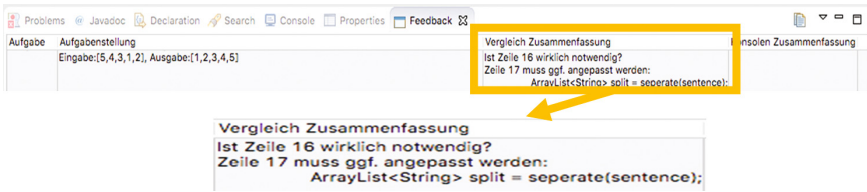


**Fig. 2.** View on feedback based on structural comparisons of learner's program and an appropriate sample program.

are the following checkboxes: (1) show insertions, (2) show deletions, (3) show replacements, and (4) show multiple. These checkboxes resemble the operations from the toolbox analysis. By selecting a checkbox, the rows in question are highlighted in different colors (depending on the corresponding operation). 'Show insertions' indicates that the highlighted code is missing in the learner solution. 'Show deletions' indicates that the highlighted code in the learner solution is possibly redundant (i.e. the line of learner's program is missing in the sample solution). 'Show replacements' operation indicates that the highlighted code area is marginally different from the sample solution (i.e. a minor alteration would transfer the learner's solution to be exactly the correspondent sample solution code line). 'Show multiple' indicates that two or more operations can be applied to the corresponding code area. Based on the findings through the toolbox, the extension retrieves lines of code from a sample program for all insertion operations. These code lines are then inserted into the learner solution and highlighted when the corresponding checkbox is selected. Additionally, highlighted code areas contain tooltip notes, which show the corresponding code line of the sample program.

### 3.3 Feedback Summary Panel

In addition to the two standalone views on feedback details, we also implemented a panel (illustrated in Fig. 3) which provides a summary of requested feedbacks and enables users to recover previous views on feedback details.



**Fig. 3.** Panel for brief summaries of previous feedback requests.

## 4 Evaluation

In September 2016, we conducted a laboratory study to investigate the impact of the prototypically implemented feedback features (as described in Sect. 3) on users' performances. Overall, 23 participants took part in the study. All participants were required to have some prior Java programming knowledge, the extent of which was assessed by a pretest. The primary focus and consideration of the study was to determine the effectiveness of the two feedback features in helping the learner develop syntactically and semantically correct programs.

### 4.1 Hypotheses

In order to evaluate the research question of whether providing feedback directly in an IDE helps learners develop correct programs, we stated the following hypotheses:

1. The extensions positively influence users' performance in comparison to those settings where users do not have any kind of help beyond Eclipse features. We, thus, expect a statistically significant difference between performances achieved when feedback features (as described in Sect. 3) are available in comparison to those performances learners achieved without using the feedback features.
2. The feedback features based on stack traces (as described in Sect. 3.1), and based on structural comparisons to sample solutions (as described in Sect. 3.2) differ in their impact on learners' performance in solving a programming task. We, thus, expect a statistically significant difference in performance between these learners who use feedback feature based on stack trace analysis in comparison to those learners who use feedback feature based on comparisons to sample programs.

Hypotheses 1 directly addresses the research question. Hypothesis 2 assesses whether EXT\_1 or EXT\_2 influences the performance of the learners to a similar or different extent.

### 4.2 Study Design

In order to measure the impact of the feedback features (as described in Sect. 3) independently from each other, we randomly assigned participants to two groups. In both groups, participants were first asked to solve a programming task (referred to as

TASK\_1) without any kind of assistance from the Eclipse IDE. Then, participants were asked to solve another programming task (referred to as TASK\_2) using the Eclipse IDE with one of the proposed feedback features. Here, we randomized the solving order of the programming tasks in order to lessen the impact of possible inequalities between the tasks. The first group could use feedback features based on stack traces (EXT\_1), and the second group could use feedback features based on structural comparisons to sample solutions (EXT\_2). As programming problems, we designed two programming tasks (referred to as FibSum80 and MergeSort, respectively). Programming task FibSum80 required participants to calculate the sum of the first 80 Fibonacci numbers. Programming task MergeSort required them to sort an integer array using the recursive merge sort algorithm. In advance, we created sample programs for both programming tasks, collecting sample solutions from multiple websites and amending them to fit the exact exercise description, splitting each sample solution into multiple steps. For each programming task, participants had 40 min for completion.

To set equal grounds among the participants regarding basic Java syntax, all participants were provided with a Java cheat sheet consisting of the basic control flow elements of Java. Participants were also provided with a hard-copy tutorial demonstrating the feedback feature (depending on the group they belonged to) and how to use it. Beyond that, they were not allowed to use any kind of help (e.g., web search engines).

For analysis purposes, we logged selected types of information while a user is using the extension. This information included the amount of extension accesses, the time spent in each extension, the source code at the time of accessing the extension, and the final programs.

### 4.3 Results

All of the learners' final programs were rated on a scale from 0 to 100%, considering three criteria. The results of a Shapiro-Wilk test of normality for each dataset for hypothesis 1 indicated that the datasets were not normally distributed. We thus used a nonparametric Wilcoxon Signed Ranks Test. This test did not reveal any statistically significant difference between the two extensions ( $p = .549$ ,  $Z = -.599$  based on positive ranks). Hence, the hypothesis could not be confirmed by this data.

**Table 2.** Table of ranks used by the Wilcoxon Signed Ranks Test. Of 13 participants, 6 experienced decreased performance while using EXT\_2, 4 did not experience alterations in performance when using EXT\_2, 3 experienced increased performance.

		N	Mean rank	Sum of ranks
Performance task 2 with extension – performance task 1 without extension	Negative ranks	6	4.58	27.50
	Positive ranks	3	5.83	17.50
	Ties	4		
	Total	13		

The dataset used for analyzing hypothesis 2 was not normally distributed. Therefore, we performed a nonparametric Mann-Whitney U Test. The results did not reveal any statistically significant difference in performance, regardless of whether participants used the extension or not ( $p = .330$ ,  $Z = -1.099$  not corrected for ties). Hence, we could not reject the null hypothesis “both samples are from the same population”.

The mean rank reveals a moderate, but statistically non-significant tendency towards decreasing performance with the use of the feedback features (see Table 2).

**Table 3.** Qualitative analysis results of the participants that have used the extension at least once.

Participant	Extension	Pre-test score	Task 2	Score task 1	Score task 2	Qualitative analysis results
1	EXT_1	4.3	FibSum80	0%	0%	No help
2	EXT_1	6.7	FibSum80	0%	100%	Little help
3	EXT_1	5.7	MergeSort	66%	33%	Little help
4	EXT_1	5.4	MergeSort	66%	33%	Little Help
5	EXT_2	3.2	FibSum80	33%	66%	No help
6	EXT_2	3.7	FibSum80	44%	100%	No help
7	EXT_2	6.3	MergeSort	100%	33%	No help
8	EXT_2	5.8	MergeSort	100%	0%	Little help
9	EXT_2	5	MergeSort	66%	0%	No help
10	EXT_2	4.2	MergeSort	33%	0%	No help
11	EXT_2	3.2	FibSum80	0%	0%	Little help
12	EXT_2	2.4	FibSum80	0%	0%	No help
13	EXT_2	4.4	MergeSort	0%	0%	No help

Measured in qualitative terms, the impact of the extension was low, as the participants’ solutions mostly disregarded the provided feedback (see Table 3). The qualitative analysis was based on the following criteria: (1) if there is no indication of transferred feedback, then result to ‘no help,’ (2) if there is some indication of transferred feedback, then result to ‘little help.’ Particularly those with a low pre-test score and very primitive learner solutions at the point of requesting feedback had difficulties using the provided feedback effectively in their solutions.

## 5 Conclusion and Future Work

The data collected from the laboratory study was, in the end, not sufficient enough to confirm the stated hypotheses. This could possibly be attributed to the relatively small number of participants, the potentially overshadowing influence of the difference between the two programming tasks, and, in the case of EXT\_2, the additional factor of the choice of sample solution steps. Concerning the choice of sample solution steps, the qualitative analysis showed that inexperienced participants accessed feedback, but did

not manage to improve their solution with the provided feedback. The analysis involved the comparison of the learner solution with the presented sample solution step. In cases where the presented sample solution had a different approach than the learner's, no transfer of feedback was noticeable. This implies that the sample solution steps the feedback was based on missed some approaches and/or may have offered too little information to be of value to a novice programmer. Here, fine-grained sample solution steps might be more beneficial for novices.

According to the survey completed by participants at the end of the study, two participants who used EXT\_2 liked the idea/concept of the extension in general. A third participant reviewing EXT\_2 liked its clear structuring and uncomplicated integration into Eclipse, as well as the fact that his code was directly available in the extension for analysis. This supports the idea that the participants, in their roles as learners, would find an IDE extension such as the one created for this project useful for learning programming. If we consider this feedback in combination with the positive overall usability rating of the extensions, we can determine the idea behind this research is valid, and that this paper makes important contributions to a largely unexplored theme. The main drawback of EXT\_2 named by participants was that the feedback was unspecific and misleading, particularly when the difference between the learner program and the sample program was considerable. The claim that the extension is nonspecific is to some extent true. The extension specifically highlights code areas it has identified as divergent in some way, but it does not offer a text-based report of the finding as further feedback. This would require the extension to conduct a far more comprehensive analysis of each finding, and possibly generate a report for a great variety of different scenarios.

For future usage, the stack trace database for EXT\_1 should be extended in this way, so that a wider range of exceptions are considered. In order to improve the usability of both extensions, it may be necessary to examine the color scheme applied, as one participant found it not intuitive. A possible new feature, which was also suggested by two participants of the study, would be feedback for the Eclipse debugger. This feedback could include the automated setting of breakpoints, feedback on the usage of the Eclipse debugger, and additional feedback on specific variable contents that could have caused errors.

**Acknowledgement.** This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under the grant “DynaFIT – Learning Dynamic Feedback in Intelligent Tutoring Systems” (PI 764/6-2).

## References

1. European Commission: Coding - the 21st century skill 6 July 2016. <https://ec.europa.eu/digital-single-market/coding-21st-century-skill>. Accessed 4 Nov 2016
2. Balanskat, A., Engelhardt, K.: Computing Our Future - Computer Programming and Coding - Priorities, School Curricula and Initiatives Across Europe. European Schoolnet, Brussels (2015)



3. The White House - Office of the Press Secretary: Remarks of President Barack Obama – State of the Union Address as Delivered, 13 January 2016. <https://www.whitehouse.gov/the-press-office/2016/01/12/remarks-president-barack-obama-%E2%80%93-prepared-delivery-state-union-address>. Accessed 04 Nov 2016
4. Zweben, S., Bizot, B.: 2015 taulbee survey: continued booming undergraduate CS enrollment; doctoral degree production dips slightly. *Comput. Res. News* **28**(5), 2–60 (2016)
5. Self, J.: The defining characteristics of intelligent tutoring systems research: ITS care, precisely. *Int. J. Artif. Intell. Educ. (IJAIED)* **10**, 350–364 (1998)
6. Nienaltowski, M., Pedroni, M., Meyer, B.: Compiler error messages: what can help novices? In: *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education*, pp. 168–172 (2008)
7. Marceau, G., Fisler, K., Krishnamurthi, S.: Mind your language: on novices’ interactions with error messages. In: *Proceedings of the 10th SIGPLAN Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward! 2011)*, pp. 3–18 (2011)
8. Hristova, M., Misra, A., Rutter, M., Mercuri, R.: Identifying and correcting Java programming errors for introductory computer science students. In: *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education (SIGCSE 2003)*, pp. 153–156 (2003)
9. Murphy, C., Kim, E., Kaiser, G., Cannon, A.: Backstop: a tool for debugging runtime errors. In: *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education (SIGCSE 2008)*, pp. 173–177 (2008)
10. Hartmann, B., MacDougall, D., Brandt, J., Klemmer, S.: What would other programmers do: suggesting solutions to error messages. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2010)*, pp. 1019–1028 (2010)
11. Flowers, T., Carver, C.A., Jackson, J.: Empowering students and building confidence in novice programmers through Gauntlet. In: *34th Annual Frontiers in Education, FIE 2004*, pp. T3H10–T3H13 (2004)
12. Holland, J., Mitrovic, A., Martin, B.: J-LATTE: a constraint-based tutor for Java. In: *Proceedings of the 17th International Conference on Computers in Education*, pp. 142–146 (2009)
13. Sykes, E.: Design, development and evaluation of the Java intelligent tutoring system. *Tech. Inst. Cogn. Learn.* **8**, 25–65 (2010)
14. Abu-Naser, S., Ahmed, A., Al-Masri, N., Deeb, A., Moshtaha, E., Abu-Lamdy, M.: An intelligent tutoring system for learning Java objects. *Int. J. Artif. Intell. Appl. (IJAIA)* **2**(2), 68–77 (2011)
15. Codeanywhere Inc.: Most Popular Desktop IDEs & Code Editors in 2014, 13 January 2015. <https://blog.codeanywhere.com/most-popular-ides-code-editors/>. Accessed 04 Nov 2016
16. Maple, S.: Java Tools and Technologies Landscape Report 2016, 14 July 2016. <http://zeroturnaround.com/rebellabs/java-tools-and-technologies-landscape-2016/>. Accessed 4 Nov 2016
17. Biradar, M.: Popularity of Programming Languages, 28 July 2015. <https://maheshbiradar.wordpress.com/2015/07/28/popularity-of-programming-language/>. Accessed 4 Nov 2016
18. Stack Exchange Inc.: Developer Survey Results 2016, March 2016. <http://stackoverflow.com/research/developer-survey-2016#technology-development-environments>. Accessed 5 Nov 2016
19. Jackson, J., Cobb, M., Carver, C.: Identifying top Java errors for novice programmers. In: *Proceedings - Frontiers in Education Conference*, p. T4C (2005)
20. Jadud, J.: A first look at novice compilation behaviour using BlueJ. *Comput. Sci. Educ.* **15**, 25–40 (2005)

21. Altadmri, A., Brown, N.: 37 million compilations: investigating novice programming mistakes in large-scale student data. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 522–527 (2015)
22. Tabanao, E., Rodrigo, M., Jadud, M.: Predicting at-risk novice Java programmers through the analysis of online protocols. In: Proceedings of the Seventh International Workshop on Computing Education Research (ICER 2011), pp. 85–92 (2011)
23. McCall, D., Kölling, M.: Meaningful categorisation of novice programmer errors. In: Frontiers in Education Conference, pp. 2589–2596 (2014)
24. Becker, B.: An effective approach to enhancing compiler error messages. In: Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE 2016), pp. 126–131 (2016)
25. Denny, P., Luxton-Reilly, A., Tempero, E.: All syntax errors are not equal. In: Proceedings of the 17th ACM Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2012), pp. 75–80 (2012)
26. Gross, S., Mokbel, B., Hammer, B., Pinkwart, N.: Learning feedback in intelligent tutoring systems. *KI - Künstliche Intell.* **29**(4), 413–418 (2015)
27. Gross, S., Pinkwart, N.: How do learners behave in help-seeking when given a choice? *Int. J. Artif. Intell. Educ.* **9112**, 600–603 (2015)
28. Mokbel, B., Gross, S., Paassen, B., Pinkwart, N., Hammer, B.: Domain-independent proximity measures in intelligent tutoring systems. In: Proceedings of the 6th International Conference on Educational Data Mining (EDM), pp. 334–335 (2013)

# Using Online Synchronous Interschool Tournaments to Boost Student Engagement and Learning in Hands-On Physics Lessons

Roberto Araya<sup>(✉)</sup>, Carlos Aguirre, Patricio Calfucura,  
and Paulina Jaure

Centro de Investigación Avanzada en Educación, Universidad de Chile,  
Periodista Mario Carrasco 75, Santiago, Chile  
roberto.araya.schulz@gmail.com

**Abstract.** We present the results from a 90-min session in which 196 fourth grade students from 13 classes from 11 schools performed a series of engaging, hands-on activities. These four activities were designed to help students understand how rockets fly. During the activities, students use skateboards, toy cars, springs, balls, and water rockets to model the physics behind a rocket launch and predict the proportion of water that will lead to maximum elevation. A subset of the classes took a post-test involving 22 basic physics questions, presented in the form of a 60-min online synchronous interschool tournament. The other subset of classes also answered the same questions online in sixty minutes, though in this case they were not presented as a tournament. The students who participated in the tournament improved significantly more than the rest. Moreover, students with weak academic performance who participated in the tournament improved the most, reducing the gap with the academically stronger students. Lessons involving hands-on experiments using skateboards, toy cars and water rockets are already highly engaging. However, this experience shows that using technology to connect schools synchronously through an online tournament is a powerful mechanism for boosting student engagement and learning in core science concepts. Furthermore, we compared learning outcomes with a previous year face-to-face interschool tournament. With the online synchronous interschool tournament, students learned twice as much as they did with the face-to-face interschool tournament.

**Keywords:** Technology-enhanced learning · Affective and motivational effects · Interschool tournaments · Collaborative learning · STEM teaching and learning

## 1 Introduction

Teachers are facing enormous challenges due to several new demands. There are new, deeper and more cognitively demanding contents and practices to teach [1]. There is also an emphasis on using and developing crosscutting concepts. In addition to this, there are new requirements to increase integration among the different science disciplines, as well as integration with mathematics and other subjects [2]. Achieving such

integration is not easy for teachers that have been educated in specific contents and trained to teach isolated subjects, with no strategies to connect with the content from other subjects. There is also increasing emphasis on implementing student-centered teaching strategies [3], teaching kids how to learn by themselves, and a hands-on approach to learning. Furthermore, there is also a demand to teach so-called 21<sup>st</sup> century skills [4]. These radically new skills are now highly valued and needed in order to create a competitive advantage, while older skills are becoming obsolete due to increased automation of the economy. This means that teaching teamwork and collaboration has become increasingly relevant, as well as developing students' interpersonal strategies, such as turn taking, social sensitivity and empathy [5]. On the other hand, all of these demands must be implemented in a completely new class environment, where students have ubiquitous access to individual mobile devices. These devices are loaded with highly attractive and addictive apps, such as messaging apps and online games. This means that the teacher is in constant competition for the students' attention. This competition inside the classroom is entirely new and incredibly powerful. Such competition simply did not exist, even just a couple of years ago. On the other hand, teachers have to teach to a more diverse population from wider cultural backgrounds and with increasing demands for social inclusion.

However, teaching practices have proven to be somewhat out of line with recent changes. Several studies of classroom practice show that almost no change has taken place over the last century [3, 6]. Teachers are used to teacher-centered strategies, and some of them can be very effective when using such strategies. For example, on the Programme for International Student Assessment (PISA), the top-performing Organization for Economic Co-operation and Development (OECD) countries use more teacher-centered teaching practices in mathematics than other countries [7]. This is a good strategy, one that can be considered as being locally optimal. Furthermore, it is a very well proven and robust solution. Leaving this local maximum requires huge leaps, and this can be risky.

In order to prepare and adapt to these challenges, teachers constantly receive methodological and technological suggestions. Recently, [8] suggest 26 effective teaching strategies. For some of them, there have been extensive empirical studies of their effect. For example, in a comprehensive study of science teaching, [9] studied 23 programs. Seven inquiry-based teaching programs using science kits did not reveal any positive outcomes in terms of student achievement in science (effect size of 0.02 standard deviations). Six programs that integrate video and computer resources and are based on cooperative learning revealed positive outcomes. Although these studies cover a wide range of strategies, none of them are based on the idea of interschool or interclass tournaments.

Using tournaments is an effective strategy that was first proposed in the seventies, albeit as an intergroup tournament within a single class [10]. However, this strategy has not been widely adopted by schools. Interclass or interschool tournaments are easier to handle for the teacher, since he/she is the coach of the whole class. Interschool tournaments are an effective engagement strategy that activates inter-group social competition and collaboration mechanisms. These social mechanisms are hardwired and were very powerful during hunter-gather tribal life [11, 12]. The hunter-gatherer brain is particularly well adapted to collaborating and learning from others in order to

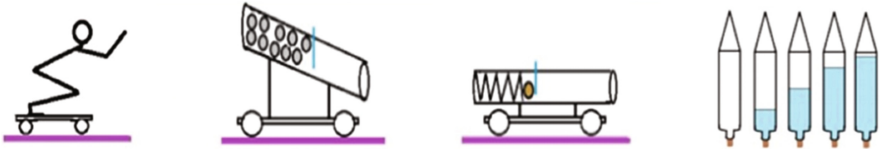
compete with neighboring groups. Cooperation is a very powerful weapon for competition. With interschool tournaments the strategy is to activate this social mechanism and the students' sense of belonging to their class or school [13–15]. Doing so boosts student engagement, collaboration, teamwork, and learning.

In this paper, we focus on a strategy based on student collaboration and social mechanisms to increase engagement. The strategy consists of online synchronous interschool tournaments. Such a strategy depends heavily on communication technology in order to link different classes and schools synchronously. The goal of this paper is to research whether existing levels of student engagement and learning with a tried and tested, hands-on lesson can be further enhanced using an online synchronous interschool tournament.

## 2 Methods

Since 2013, thirteen low Socio Economic Status (SES) schools from two districts in Santiago, Chile, have been developing a hands-on activity in fourth grade. This activity is carried out in a lesson based on what is known as “modeling instruction” [16], where students have to learn to use, adapt and build models. In this particular lesson, students do experiments using different models in order to understand how rockets fly. The students study a sequence of four experiments during a 90-min lesson. The main goals of the lesson are to help the students understand basic notions of physics, such as motion and forces, as well as some basic scientific practices, such as modeling, making predictions with models, using empirical measurements to validate or reject a model, explaining the results and carrying out a peer review process with the written explanations. One of the experiments involves jumping from a skateboard (Fig. 1). The students experiment by jumping under different load conditions, such as having different loads in their backpacks. This experiment is a modeling instruction activity, where the jumping student models the water, while the skateboard models the rocket. This experiment is an initial hands-on activity that introduces the students to the concept of the conservation of momentum. A second experiment involves balls and a toy car used as a cart, as shown in Fig. 1. This is another way of modelling water rockets, where the balls model the water, and the cart models the rocket. This experiment provides the students with a second view of the concept of the conservation of momentum. The third experiment involves a cart with a spring and a ball, as shown in Fig. 1. Students experiment using different numbers of balls, while reflecting on how this is similar to the water rocket. In this case, the cart models the rocket, the ball models the water, and the spring models the air pressure inside the plastic bottle (the rocket). The fourth experiment consists in throwing the bottle after pumping air to a certain pressure. This is done in 5 different conditions. Each condition corresponds to a fraction of water, which goes from without water, a quarter of water, half of water, three quarters of water, to the bottle filled with water. Each of the first 3 experiments takes about 15 min, but the last lasts 25 min.

Every year, the students take a pre-test and an identical post-test comprising 15 multiple-choice questions that look to measure conceptual understanding of each of the four experiments. Most questions have 5 options but some of them have 3, others 4 and



**Fig. 1.** Four types of question on the pre-test and post-test. For the first type of question, the students have to say in which direction and at what velocity will the skateboard travel. For the second type of question, they have to say where the cart will go and at what velocity when the balls are released under different conditions (i.e. number of balls). The third type of question asks the students where the cart will move when the spring is released under different conditions. Finally, the fourth question is about water rockets, such as the proportion of water that is needed in order to reach maximum elevation.

others 6 or more. Each question is graded on a scale of 1 (minimum) to 7 (maximum). This is the standard scale used in Chile. For questions 1 to 15, answering at random produces an average score of 2.1, whereas for questions 16 to 22 random answers lead to an average score of 2.2. Student performance on the pre-test has been only slightly above the results obtained by answering at random, although the difference is statistically significant. There are some questions with intuitive answers, where the level of the students' responses was well above the random score. However, other questions are more difficult. Therefore, the students' performance on these questions has traditionally been below the random score.

In 2016, there was a change in the activity, with the introduction of an online synchronous interschool tournament. The tournament was conducted using a STEM platform in the cloud. This is a platform where the classroom teacher and a remote teacher track student performance in real time, detect which students are having difficulty, and provide just-in-time support using a chat function included in the platform. This platform has the ability to synchronize whole courses from different schools to perform online and synchronous tournaments between schools.. In this paper, we analyze the effect of this technological intervention. The activity was held in November and December of 2016. A total of 367 fourth grade students took the pre-test during a single session. In the following 90-minute session, the students then went to a science lab to take part in the experimental lesson. This lesson included four experiments. During December, 239 students took the post-test. Several schools could not take the post-test during December due to time constraints that are typical at the end of the school year. In total, 196 students took both the pre-test and post-test. The statistics shown in this paper will be restricted to these 196 students. Of the 196 students, 88 were girls and 108 were boys. For the purpose of our analysis, we also classify students according to their academic performance. In order to do so, we use their Grade Point Average (GPA) for the year in science and math. Those with a GPA below their class average are classified as academically weak students, while the rest are considered academically strong. Therefore, using this classification method there are 96 academically weak and 100 academically strong students.

Not all schools could participate in the online synchronous interschool tournament. Given the time and scheduling restrictions, only three classes were able to participate.

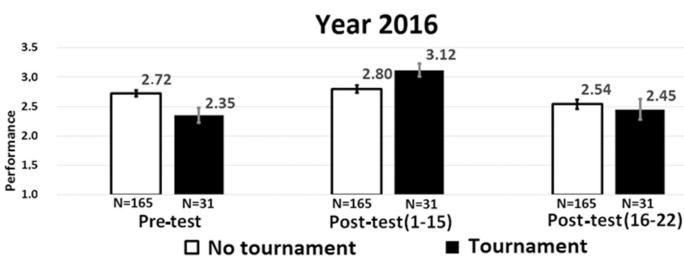
Therefore, 31 students from these three classes participated in the online synchronous interschool tournament, while 165 students answered the same questions online using the same platform, although not in tournament mode.

It is important to note that during the pre-test and the 4 experiments conducted in the science lab the students knew that there was a tournament. However, they did not know whether they were going to be able to participate in the online synchronous interschool tournament. The classroom teachers also did not know whether their students would participate in the tournament. Nevertheless, all of them knew that their average performance was going to be automatically published on the online platform and listed alongside the average performance of the other classes. Throughout the year, the class' average performance in every session is automatically published on the online platform. Furthermore, it is listed alongside the average performance of the classes from the other schools.

As mentioned previously, this exact same lesson has been taught in the same schools since 2013. The only difference in this case was the inclusion of the online synchronous interschool tournament. In 2013, students did not participate in an online synchronous interschool tournament. However, they did participate in a face-to-face tournament that took place in the municipal gym. A total of 204 students participated that year and took the same pre-test and post-test as the students in 2016. The post-test was taken a couple of days before the tournament. We will therefore also compare the effects of both types of tournament.

### 3 Results

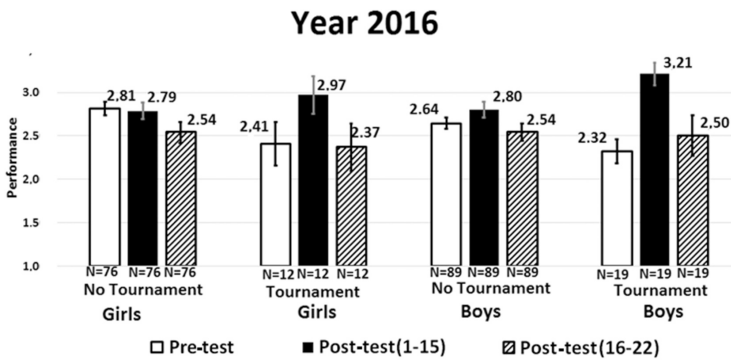
First, we analyze the results from 2016. As shown in Fig. 2, the 165 students that did not participate in the tournament (no tournament condition) performed only slightly better on the post-test when compared with the pre-test. In that sense, the learning is moderate. The effect size is 0.11 standard deviations and is statistically significant. On the other hand, the 31 students that participated in the online synchronous interschool tournament (tournament condition) enjoyed a high level of improvement. The effect size is 1.09 standard deviations, and it is also statistically significant. It is important to



**Fig. 2.** Performance on the 15 questions on the pre-test and post-test, as well as the 7 extra questions on the post-test, for both the group of 167 students that did not participate in the tournament, as well as the 31 students that did participate in the online-synchronous interschool tournament.

note that the students who participated in the tournament did much worse on the pre-test than the other students. However, on the post-test they did much better. Nevertheless, on the 7 completely new questions (questions 16 to 22), which measure generalization and a deeper conceptual understanding, both groups performed similarly. Although there is a small difference, it is not statistically significant. Given the poorer performance on the pre-test by the students who participated in the tournament, the fact that the students performed similarly on questions 16 to 22 suggests that the improvement was greater for the online synchronous tournament group.

It is interesting to compare learning according to gender. As shown in Fig. 3, girls in the no tournament condition did not improve. The effect size is  $-0.02$ . However, in the tournament condition the effect size was  $0.65$ . On the other hand, boys in the no tournament condition improved and the effect size was  $0.25$  standard deviations. However, boys in the tournament condition enjoyed a marked improvement. In this case, the effect size was  $1.48$  standard deviations.

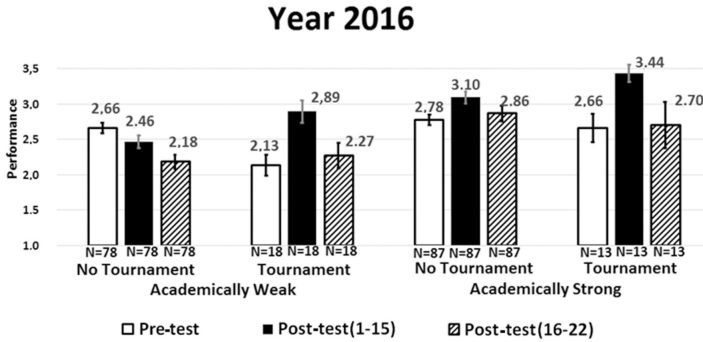


**Fig. 3.** Performance by boys and girls on the 15 questions included on the pre-test and post-test, as well as on the 7 extra questions on the post-test, for both the group of 167 students that did not participate in the tournament, as well as the 31 students that did participate in the online synchronous interschool tournament.

Let us consider the learning outcomes of the academically weak and academically strong students. As shown in Fig. 4, the academically weak students did not improve in the no tournament condition. However, they improved considerably in the tournament condition. In this case, the effect size was  $1.24$  standard deviations. On the other hand, the academically strong students improved in both conditions. The effect size for the no tournament students was  $0.47$  standard deviations, while for the tournament students the effect size was  $1.06$  standard deviations. It is also very interesting to note that even though the academically weak students in the tournament condition performed much worse on the pre-test than the academically weak students in the no tournament condition, they performed slightly better on the generalization questions (questions 16–22). Although the difference is not statistically significant, the tournament condition made it possible to close the gap that was present on the pre-test.

Now, we will compare the results obtained in 2016 with the results obtained in 2013. Although they are different students, all of them were fourth graders from the





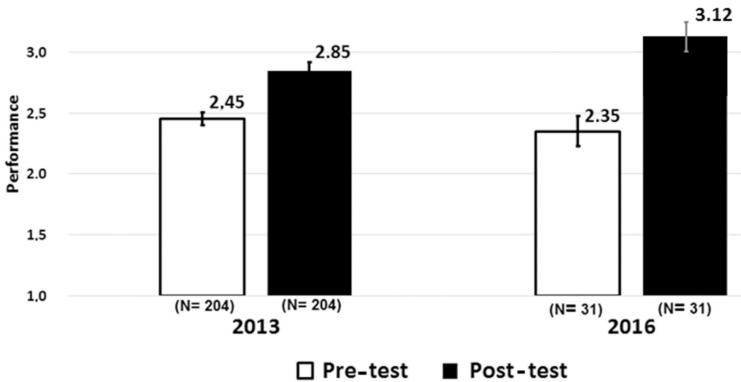
**Fig. 4.** Performance by academically weak and academically strong students on the 15 questions on the pre-test and post-test, as well as on the 7 extra questions on the post-test, for both the group of 167 students that did not participate in the tournament, as well as for the 31 students that did participate in the online synchronous interschool tournament.

same schools. As mentioned previously, the lesson involved the same four experiments. This lesson has been taught at the same schools since 2013. However, in 2013 a face-to-face tournament was organized instead. A couple of days after taking the post-test, all of the classes met at the municipal gym. These classes participated in a tournament based on launching water rockets. However, the score for each school depended heavily on the post-test. All of the students knew this was the case several days before taking the post-test. This was a highly attractive activity for the students. However, the logistics of organizing the event were not easy, as 13 classes from 11 schools had to be transported at the same time from their respective schools to the municipal gym. Moreover, according to the organizers and teachers, the whole activity was very demanding and exhausting. Therefore, the superintendent decided to cancel the tournament in 2014 and 2015. However, in 2016 the tournament was replaced by an online synchronous tournament. The plan was to connect schools using technology and therefore avoid transporting the classes. It is therefore very important to compare the effect of the face-to-face tournament in the municipal gym with the online synchronous tournament. Figure 5 shows the effect size of both tournaments. In 2013, the face-to-face tournament had an effect size of 0.51 standard deviations, whereas in 2016 the online synchronous tournament had an effect size of 1.10 standard deviations.

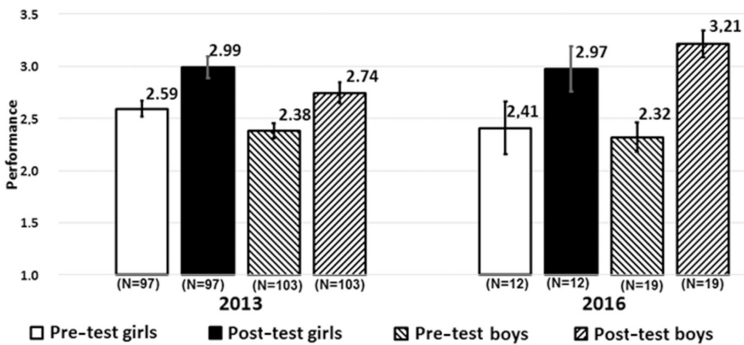
Therefore, the effect size of the online synchronous interschool tournament is more than twice the effect size of the face-to-face tournament. This really is a completely unexpected finding. It shows that technology cannot only solve a logistical problem of a very engaging educational activity; it can also produce highly impressive improvements in learning outcomes.

Let us now analyze how the type of tournament impacts the learning outcomes according to gender and the students' academic performance.

As shown in Fig. 6, the effect size for girls in the 2013 face-to-face tournament is 0.54 standard deviations, whereas for the 2016 online synchronous interschool tournament it was 0.65 standard deviations. This is an increase of 20% in the effect size. In the case of boys, the increase in the effect is greater. In the 2013 face-to-face



**Fig. 5.** Performance on the 15 questions on the pre-test and post-test, as well as at the 2013 face-to-face tournament and on the online synchronous tournament held in 2016.



**Fig. 6.** Performance on the 15 questions on the pre-test and post-test, as well as at the face-to-face tournament in 2013 and on the online synchronous tournament in 2016.

tournament, the effect size is 0.50 standard deviations, whereas in the 2016 online synchronous interschool tournament it was 1.48 standard deviations. This is an increase of 196% in the effect size. This is a huge increase. Given that in the face-to-face tournaments the learning outcomes are similar among boys and girls and that this is not the case with online synchronous tournaments, it seems that the networking technology has a significant effect on boys for tournaments; much higher than the effect on girls.

Figure 7 reveals the performance by academically weak and academically strong students for both tournaments. The effect size for academically weak students involved in the 2013 face-to-face tournament is 0.22 standard deviations, whereas in the case of the 2016 online synchronous interschool tournament it was 1.24 standard deviations. This is an increase of 464% in the effect size. This is a huge increase in effect size. On the other hand, for the academically strong students who participated in the 2013 face-to-face tournament the effect size was 1.00 standard deviations, whereas for the 2016 online synchronous interschool tournament it was 1.06 standard deviations. This means that the effect sizes are similar.

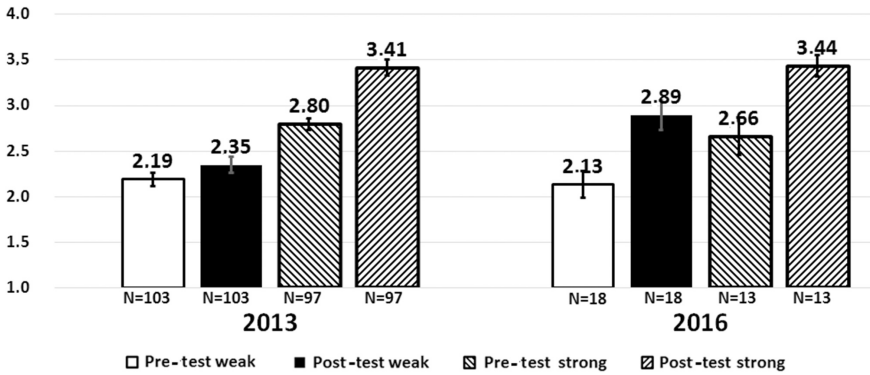


Fig. 7. Performance on the 15 questions on the pre-test and post-test, as well as at the face-to-face tournament in 2013 and the online synchronous tournament in 2016.

### 4 Conclusions

STEM teaching and learning is a huge challenge. It requires a significant change in teaching practices in order to teach new crosscutting concepts, increase integration across subjects; all within the context of a classroom where mobiles devices are ubiquitously present and constantly competing to attract the students’ attention. Several powerful teaching strategies have been suggested. For example, [8] propose 26 “scientifically proven approaches”. These include hands-on activities, making or producing practical knowledge, student participation, asking for self-explanations and undoing misconceptions. Hands-on experimental work has been studied in particular [17]. This leads to highly engaging and meaningful activities. Another proposed strategy is to increase the level of excitement, such as in games [18], particularly social games. Games activate the mechanism of social facilitation [15], where people and other animals perform better when other subjects are around [19, 20]. Another important strategy is to activate the mechanism involved in the sense of belonging. This is the sense of being accepted, valued, and included. According to [21] “belongingness appears to have multiple and strong effects on emotional patterns and on cognitive processes” p. 407. This is a great motivational resource available in every school and class. Interestingly, the list [8] of 26 strategies does not include the strategy of running interclass or interschool tournaments. This is a powerful strategy for activating social mechanisms, such as a sense of belonging, teamwork, engagement, and excitement. Moreover, technology can make a big different to implementing this strategy.

Data collected from these tournaments show that online and synchronous tournaments between courses have a greater effect on boys than girls. This effect is consistent with other results from evolutionary psychology. According to Geary [22, 23], boys tend to form larger groups, which is normal when preparing for inter-tribal conflicts. Girls instead tend to form much smaller groups, with more intense and lasting relations. Thus, boys are more easily motivated by large group collaboration in preparation for

inter-group conflicts. Therefore, a prediction of evolutionary psychology is that the use of the competition mechanism between courses should motivate more boys than girls to prepare and pay attention in the tournament.

There is an extensive literature on the importance of play for learning both in humans and in other animals [24]. In addition, since at least the decade of the 70, there are studies of the important effects on learning and attitudes of the competitions and tournaments between teams [25, 26]. However, to the best of our knowledge there are two questions that have not been addressed before. Firstly, with the use of technology, do online synchronous interschool tournaments enhance hands-on experimental sessions that are already engaging, such as the ones involving water rockets? The second question asks whether the learning outcomes obtained with online synchronous interschool tournaments is similar to or better than those obtained with face-to-face tournaments, where students from several schools meet at the same physical place in order to participate in an interschool competition? These tournaments are run using strategies from TV shows in order to increase engagement. Students wear school uniforms and display their school flags in order to enhance the sense of belonging. Music and school songs activate the tribal mechanism of intergroup competition. It is not clear whether technology-enhanced tournaments can trigger the same emotional mechanisms and produce the same level of engagement. However, our results from years of interschool tournaments give some preliminary empirical evidence to suggest that technology makes possible to implement online synchronous interschool tournaments that can make an important difference to learning STEM in elementary schools. This finding is very interesting and we do not yet have a definitive explanation. One possibility is that the online and synchronous tournaments between courses, allow to individualize and account with much more precision the contribution of each player. Although the STEM platform periodically announces only the team's score, it nevertheless constantly appoints the student by name and gives him/her feedback on his individual performance and congratulates him personally for his/her successes. That's impossible to do with hundreds of students playing in a face-to-face tournament in a gym. This is a facility that provides technology for online and synchronous tournaments, and offers a great advantage over face-to-face tournaments. In the near future we hope to study this impact in more depth, and also include other STEM contents.

**Acknowledgments.** Funding from PIA-CONICYT Basal Funds for Centers of Excellence Project FB0003 is gratefully acknowledged and to the Fondef D15I10017 grant from CONICYT.

## References

1. Quinn, H., Schweingruber, H., Keller, T. (eds.): *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. National Academies Press, Washington, D.C. (2012)
2. Honey, M., Pearson, G., Schweingruber, H. (eds.): *STEM Integration in K-12 Education: Status, Prospects, and An Agenda for Research*. National Academies Press, Washington, D.C. (2014)
3. Cuban, L.: *Inside the Black Box of Classroom Practice Change Without Reform in American Education*. Harvard Education Press, Cambridge (2013)

4. Avent, R.: *The Wealth of Humans. Work, Power, and Status in the Twenty-First Century.* St. Martin's Press, New York (2016)
5. Colvin, G.: *Humans are Underrated What High achievers Know That Brilliant Machines Never Will.* Penguin, London (2016)
6. Labaree, D.: *Someone Has to Fail.* Harvard University Press, Cambridge (2010)
7. OECD, *Ten Questions for Mathematics Teachers...and How PISA Can Help Answer Them* (2016)
8. Schwartz, D., Trang, J., Blair, K.: *The ABCs of How we Learn.* Norton & Company Inc., New York (2016)
9. Slavin, R., Lake, C., Hanley, P., Thurston, A.: Experimental evaluations of elementary science programs: a best-evidence synthesis. *J. Res. Sci. Teach.* **51**(7), 870–901 (2014)
10. Edwards, K., De Vries, D.; Snyder, J.: *Games and teams: a winning combination.* Report 135. Center for Social Organization of Schools. Johns Hopkins University (1972)
11. Henrich, J.: *The Secret of Our Success.* Princeton University Press, Princeton (2016)
12. Greene, J.: *Moral Tribes.* The Penguin Press, New York (2013)
13. Araya, R., Jimenez, A., Bahamondez, M., Dartnell, P., Soto-Andrade, J., González, P., Calfucura, P.: Strategies used by students on a massively multiplayer online mathematics game In: *Advances in Web-based Learning - ICWL 2011.* LNCS, vol. 7048 (2011)
14. Araya, R., Jimenez, A., Bahamondez, M., Dartnell, P., Soto-Andrade, J., Calfucura, P.: Teaching modeling skills using a massively multiplayer on line mathematics game. *World Wide Web J.* **17**(2), 213–227 (2014). Springer Verlag
15. Araya, R., Aguirre, C., Bahamondez, M., Calfucura, P., Jaure, P.: Social facilitation due to online inter-classrooms tournaments In: LNCS, vol. 9891, pp. 16–29 (2016)
16. Perazoo, J., Pehhyacker, C., Stronck, D., Bass, K., Heredia, J., Lobo, R., Ben-Shalom, G.: Persistent and encouraging achievement gains on math common core items for middle school english language learners. *Sci. Educ. Civ. Engagem.* **7**(2) (2015)
17. Abrahams, I., Millar, R.: Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *Int. J. Sci. Educ.* **30**(14), 1945–1969 (2008)
18. Devlin, K.: *Mathematics Education for a New Era: Video Games as a Medium for Learning.* Peters/CRC Press, Boca Raton (2011)
19. Zajonc, R.: Social facilitation. *Science* **149**, 269–274 (1965)
20. Zajonc, R.: Attitudinal effect of mere exposure. *J. Pers. Soc. Psychol.* (1968)
21. Baumesiter, R., Leary, M.: The need to belong: desire for interpersonal attachment as a fundamental human motivation. *Psychol. Bull.* **117**(3), 407–529 (1995)
22. Geary, D., Byrd-Craven, J., Hoard, M., Vigil, J., Numtee, C.: Evolution and development of boys social behavior. *Dev. Rev.* **23**, 44–470 (2003)
23. Geary, D.: Educating the evolved mind: conceptual foundations for an evolutionary educational psychology. In: Carlson, J.S., Levin, J.R., Greenwich, C.T. (eds.) *Psychological Perspectives on Contemporary Educational Issues*, vol. 42. Information Age Publishing, Charlotte (2007)
24. Pellegrini, A.: *The Role of Play in Human Development.* Oxford University Press, New York (2009)
25. Edwards, K., DeVries D.: The effects of teams-games-tournaments and two structural variations on classroom process, student attitudes, and student achievement (Technical report 172). Baltimore: Johns Hopkins University, Center for Social Organization of Schools (1974)
26. Hulten, B., DeVries, D.: Team competition and group practice: Effects on student achievement and attitudes (Technical report 212). Baltimore: Johns Hopkins University Center for Social Organization of Schools (1976)

# Story-Based Multimedia Analysis Using Social Network Technique

Quang Hai Bang Tran<sup>1,2</sup>, Thi Hai Binh Nguyen<sup>2</sup>, Phong Nha Tran<sup>2</sup>,  
Thi Thanh Nga Tran<sup>3</sup>, and Quang Dieu Tran<sup>2</sup>(✉)

<sup>1</sup> University of Information and Technology, Ho Chi Minh City, Vietnam  
bangtqh.ncs@grad.uit.edu.vn

<sup>2</sup> University of Transport and Communications, Ho Chi Minh City, Vietnam  
{bangtqh,binhnh,tpnha,dieutq}@utc2.edu.vn

<sup>3</sup> Nong Lam University, Linh Trung, Thu Duc District,  
Ho Chi Minh City, Vietnam  
ngattt@hcmuaf.edu.vn

**Abstract.** Recent years, the number of multimedia contents is increasing rapidly. The demand of proposing an efficient method for analyzing a multimedia content has becoming a hot research topic. Many proposed methods have introduced including low-level processing, structured analysis, story-based analysis and so on. However, such methods have some unsatisfactory results because of speed, time processing. In this paper, we provide a new method in analyzing the story of a multimedia content by using social network analysis techniques. The experimental results showed that our method is able to discover storytelling of the given multimedia content with good accuracy performance and processing.

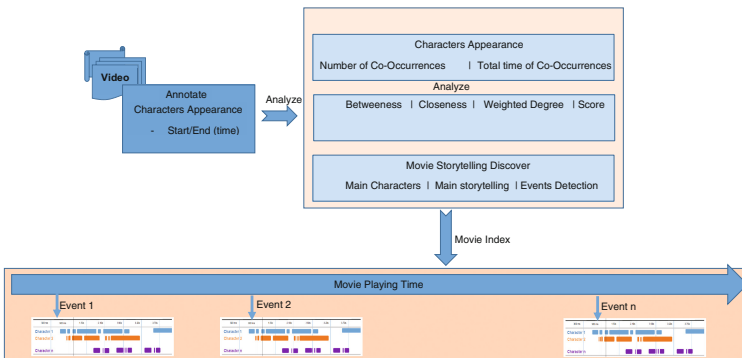
**Keywords:** Social network analysis · Story-based multimedia analysis · Multimedia analysis

## 1 Introduction

Nowadays, due to the raising of the number of multimedia contents. Proposing an effective method to analyze and explore the story of a multimedia content has become a hot research topic in recent researches, in which story-based multimedia analysis is a method that takes into account discovering hidden information from the given multimedia content. Such methods have been focused in internal or external features to determine and discover objects, text and so on by using features extraction and detection. While content-based methods are founded on low-level syntactic characteristics of individual multimedia content objects such as color, texture, shape, and motion. But these techniques might not work well for detecting physical commonalities among frames, they have a failure short in correlating with semantic contents of frames. Story-based are focused on the hidden information in a multimedia content. These approaches aim to help users to produce a shortened version by removing unimportant or redundant contents

of the multimedia content. The proposed methods generally used film and drama theories to analyze plot-lines and index the movie by mainly focusing on characters and relationships among them [3]. In this regards, such methods have been overcome this issue by using events extraction and objects recognition. For example, Li et al. [4] provided a concept in analyzing and recognizing events in the dialogs of characters. In the study of Weng et al. [9], the recognition technique had been applied for recognizing characters in the given movie. In this regards, these methods have some limitations, including the accuracy rate of face detection/recognition and audio analysis. However, these studies could not detect the protagonist or how the event important is. Recent research on movie understanding has focused on the audience [8], story-based and content-based analysis [4,9]. These approaches aim to help the user to produce a shortened version by removing unimportant or redundant content in the movie.

In this paper, we propose a method to represent a multimedia content based on the occurrences and co-occurrences of the characters using time temporal characteristics. From these analyzing results, we used social networks technique to analyze the characters time aspect of the given multimedia content. For the proposed method, we build a model for the given multimedia content as temporal characteristics. Then, the social network is used to classify the characters in certain communities, to discover main character(s) and events in the given multimedia content. The analyzing of character network and events detection then used to extract index from the given multimedia content. Figure 1 illustrates the framework of this proposed method.



**Fig. 1.** Multimedia content indexing framework

The rest of this paper is organized as the following. In Sect. 2, we propose a new method for presenting the occurrences and co-occurrences of the characters in a multimedia content. The analyzing of story-based is presented in Sect. 3. The evaluation results is examined in Sect. 4. The conclusion and future works is presented in Sect. 5.

## 2 The Occurrences of Characters

### 2.1 Character's Occurrences Model

### 2.2 Representation of Characters' Appearance Aspects

A character in a multimedia content has a time distribution. We can represent the character's playing time distribution by means of a sequence of time intervals as the following.

Let  $C = \{c_1, c_2, \dots, c_k\}$  be the set of characters in a multimedia content.  $MTL$  is the total length of the multimedia content that is calculated by a timestamps value. Time intervals is represented by  $[t_1, t_2]$  where  $t_1$  and  $t_2$  are timestamps, and  $t_2 > t_1$ .

**Definition 1 (Character's Appearance Distribution).** *Let  $c_i \in C$  be a character in a multimedia content. Character's appearance distribution of a character  $c_i$  in the given multimedia content is represented as*

$$\begin{aligned} T(c_i) &= \langle t_1^i, t_2^i, \dots, t_n^i \rangle \\ &= \langle [t_1^{i-}, t_1^{i+}], \dots, [t_n^{i-}, t_n^{i+}] \rangle \end{aligned} \quad (1)$$

where  $c_i \in C$ ;  $t_j^{i+} > t_j^{i-}$  for  $i = 1, \dots, k$ ;  $j = 1, \dots, n$ ;  $t_j^i \in [0..MTL]$ ; in a time point not belonging to intervals  $t_1^i, \dots, t_n^i$  character  $c_i$  does not appear.

**Definition 2 (Character's Disappearance Distribution).** *Character's disappearance distribution is the time that the character does not appear in the multimedia content. Let  $c_i \in C$  be a character in the multimedia content, this distribution is represented as*

$$\begin{aligned} T'(c_i) &= \langle t_1'^i, t_2'^i, \dots, t_n'^i \rangle \\ &= \langle \{t_2^{i-} - t_1^{i+}\}, \dots, \{t_n^{i-} - t_{n-1}^{i+}\} \rangle \end{aligned} \quad (2)$$

where  $t_j^{i-} - t_{j-1}^{i+} > 0$  for  $i = 1, \dots, k$ ;  $j = 1, \dots, n$ ;  $t_j^i \in [0..MTL]$ .

**Definition 3 (Total of Time Intervals).** *Let  $T = \{t_1, t_2, \dots, t_m\}$  be a character's appearance distribution, denote  $|T|$  is total of time intervals of  $T$ .  $|T|$  can determine as*

$$|T| = \bigcup_1^m (t_i) : i = 1..m \quad (3)$$

Suppose that we have two character's appearance distributions  $T(c_i)$  and  $T(c_j)$ . We can determine the time intervals in which they appear together and owing to this we can determine total time and number of their co-occurrence.

**Definition 4 (Co-occurrence Time Distribution).** *Suppose that  $c_i, c_j \in C$  are two characters in a multimedia content and  $T(c_i) = \langle t_1^i, t_2^i, \dots, t_n^i \rangle$ ,  $T(c_j) = \langle t_1^j, t_2^j, \dots, t_m^j \rangle$ , are two character's appearance distributions. Co-appearance Time Distribution of two characters  $c_i$  and  $c_j$  is defined as*

$$T(c_i \wedge c_j) = \langle t_p^i \cap t_q^j : t_p^i \in T(c_i), t_q^j \in T(c_j) \rangle \quad (4)$$



**Definition 5 (Total Co-occurrence Time of Characters).** *Total time of characters co-occurrence  $\alpha_{ij}$  is the total of intervals in  $T(c_i \wedge c_j)$ , where  $T(c_i \wedge c_j)$  is character's co-occurrence time distribution of characters  $c_i$  and  $c_j$ .*

$$\alpha_{ij} = |T(c_i \wedge c_j)| \quad (5)$$

where  $T(c_i \wedge c_j)$  is character's co-occurrence time distribution characters of  $c_i$  and  $c_j$ .

We also represent the number of co-occurrence between characters as follows.

**Definition 6 (Number of Co-occurrences Time of Characters).** *Number of characters co-occurrences  $\beta_{ij}$  is the number of intervals in  $T(c_i \wedge c_j)$ , where  $T(c_i \wedge c_j)$  is character's co-occurrence time distribution of the characters  $c_i$  and  $c_j$ .*

We can note that  $\alpha_{ij=0}$  iff  $\beta_{ij} = 0$ .

### 3 Story-Based Multimedia Analysis

To analyze relationships among characters in a multimedia content, we define a graph as the following.

**Definition 7 (Character Network).** *The character network using characters time aspect is a undirected-weight graph describes as*

$$G = \langle C, R \rangle$$

where  $C = \{c_1, c_2, \dots, c_k\}$  is the set of characters in a multimedia content;  $R = \{(c_i, c_j, (\alpha_{ij}, \beta_{ij})) : c_i, c_j \in C, \alpha_{ij} > 0 \text{ or } \beta_{ij} > 0\}$

We can note that if  $(c_i, c_j, (\alpha_{ij}, \beta_{ij})) \in R$  iff  $(c_j, c_i, (\alpha_{ji}, \beta_{ji})) \in R$ .

The most important task of multimedia content analysis is to determine main character(s) and explore multimedia content storytellings. In this work, we apply Betweenness Centrality, Closeness Centrality and Weighted Degree to analysis character network and define a social score of characters based on integrated Betweenness Centrality, Closeness Centrality and Weighted Degree measurement values as the following.

**Definition 8 (Score of Character).** *The social score of a character in the multimedia content is defined as*

$$S_{c_i} = \frac{BC(c_i) + CC(c_i) + WD(c_i)}{3} \quad (6)$$

where  $c_i$  is  $i^{th}$  character,  $S(c_i)$  is the score of appearance value,  $BC(c_i)$  is the Betweenness Centrality value,  $CC(c_i)$  is the Closeness Centrality value,  $WD(c_i)$  is the Weighed Degree value in the multimedia content.

**Definition 9 (Main Character).** A character  $c_i$  is the protagonist of a multimedia content if  $S_{c_i}$  is the largest.

**Definition 10 (Significant Class - Ma).** Significant class of a multimedia content is a class, in which, contains the characters, while it tell the main story of the multimedia content. Significant class of the multimedia content is defined as:

$$Ma = \{c_i : S_{c_i} \geq \frac{1}{k} \sum (S(c_p))\} \quad (7)$$

where  $c_i \in C$  is a character in a multimedia content;  $S_{c_i}$  is the social score of  $c_i$ ;  $p = [1..k]$ .

**Definition 11 (Minor Class - Mi).** Minor class of a multimedia content is the class, in which, contains the characters, while it not belong to significant class. Minor class of the multimedia content is defined as

$$Mi = \{c_i : S_{c_i} < \frac{1}{k} \sum (S(c_p))\} \quad (8)$$

where  $c_i \in C$  is a character in the multimedia content;  $S_{c_i}$  is the social score of  $c_i$ ;  $p = [1..k]$ .

Algorithm 1 illustrates the algorithm for building character network. We calculate and build it based on the total time of characters co-occurrence and the number of characters co-occurrences, then its nodes will be measured by using social network centralities to extract main characters, protagonist and classify them into certain groups.

**Definition 12 (Events in a Multimedia Content).** Let  $C = \{c_1, c_2, \dots, c_k\}$  be the set of characters a multimedia content and  $MTL$  is the length of a the multimedia content.  $T = T(c_1), T(c_2), \dots, T(c_k)$  be the set of character's occurrences distribution of characters in the multimedia content. Events  $E$  is a sequence that is defined as

$$E = \langle C_u[t^-, t^+] : 0 < t^- < t^+ < MTL, C_u \subseteq C \rangle. \quad (9)$$

We note that  $c_i[t_p^{i-}, t_p^{i+}] \in E$  iff  $c_i \in C_u : [t_p^{i-}, t_p^{i+}] \subseteq [t^-, t^+]; 0 \leq p \leq MTL$ .

Suppose that  $c_u[t_p^{u-}, t_p^{u+}]$  and  $c_v[t_r^{v-}, t_r^{v+}]$  are two time intervals of  $c_u$  and  $c_v$  of a multimedia content.  $e_v \in E$  can be calculate as

$$e_v = c_u[t_p^{u-}, t_p^{u+}] \cup c_v[t_r^{v-}, t_r^{v+}] \quad (10)$$

where  $0 < t_r^{v-} < t_p^{u+}; [t_p^{u-}, t_p^{u+}] \cap [t_r^{v-}, t_r^{v+}] > 0$ .

In a multimedia content, an event contain in itself many useful information. For our proposed method, we classify events into three groups: main event - including the event, in which, contains main character(s); sub event - in which contain minor characters and hybrid event - in which include both of main and minor characters. By using Eqs. 9 and 10, we need to evaluate each  $e_v \in E$  and

**Algorithm 1.** Character Network Algorithm

---

Let  $C = \{c_1, c_2, \dots, c_k\}$  be the set of characters in a multimedia content.  
Let  $T(c_i \wedge c_j)$  is Co-Occurrence Time Distribution of a multimedia content.

```

procedure CHARACTER-NETWORK
  while  $t \in T(c_i \wedge c_j)$  do
    Calculate  $\alpha_{ij}$ 
    Calculate  $\beta_{ij}$ 
  end while
end procedure
procedure CENTRALITY-NODE
  for each node  $c_i \in C$  do
    Calculate  $BC(c_i)$ 
    Calculate  $CC(c_i)$ 
    Calculate  $WD(c_i)$ 
    Calculate  $S(c_i) = \frac{1}{3}(BC(c_i) + CC(c_i) + WD(c_i))$ 
  end for
end procedure
procedure CHARACTER-CLASSIFICATION
  for each node  $c_i \in C$  do
    classify  $c_i$  into  $Ma$  or  $Mi$ 
  end for
  return  $c_i : S(c_i) = \max(S(c))$ 
end procedure

```

---

classify it in to certain group of events. These group then use for multimedia content indexing.

Let  $E$  be the set of events in a multimedia content. We classify  $E$  by three classes of event as the following.

**Definition 13 (Main Events).** *Main events class of a multimedia content is a class, in which, each event contains the appearance of main characters. Main events of the multimedia content is defined as*

$$E_{main} = \{e_v : (c_v \in e_v) \in Ma\} \quad (11)$$

**Definition 14 (Sub Events).** *Sub events class of a multimedia content is a class, in which, each event contains the appearance of minor characters. Sub events class of the multimedia content is defined as*

$$E_{sub} = \{e_v : (c_v \in e_v) \in Mi\} \quad (12)$$

**Definition 15 (Hybrid Events).** *Hybrid events class of a multimedia content is a class, in which, each event contains the appearance of main and minor characters. Hybrid events class of the multimedia content is defined as*

$$E_{hyb} = \{e_v : (c_v \in e_v) \in Mi \text{ or } (c_v \in e_v) \in Ma\} \quad (13)$$

The algorithm for events detection and multimedia content indexing is illustrated as the Algorithm 2. The goals of this algorithm is classify event in the

**Algorithm 2.** Movie Indexing Algorithm

---

Let  $\varepsilon$  be the average of the characters' time silent  
 Let  $c_u[t_p^{u-}, t_p^{u+}], c_v[t_r^{v-}, t_r^{v+}] \in C$  be two characters appearance distribution of  $c_u$  and  $c_v$  of a multimedia content.  
 Let  $E$  be the set of events in a multimedia content.

**procedure** EVENT DETECTION  
   **for**  $c_u[t_p^{u-}, t_p^{u+}], c_v[t_r^{v-}, t_r^{v+}] \in T(c)$  **do**  
      $e_v = c_u[t_p^{u-}, t_p^{u+}] \cup c_v[t_r^{v-}, t_r^{v+}] : t_r^{v-} - t_p^{u+} < \varepsilon$   
      $E = E + \{e_v\}$   
   **end for**  
**end procedure**

**procedure** MOVIE INDEXING  
   **for** each node  $c_v \in C$  **do**  
      $E_{main} = \{e_v : (c_v \in e_v) \in Ma\}$   
      $E_{sub} = \{e_v : (c_v \in e_v) \in Mi\}$   
      $E_{hyb} = \{e_v : (c_v \in e_v) \in Ma \text{ or } e_v : (c_v \in e_v) \in Mi\}$   
   **end for**  
**end procedure**

---

multimedia content as the indexing. Figure 5 gives an example of indexed multimedia content. As shown, the multimedia content “*Empire Strikes Back*” is first decomposed into a series of events, where each event has certain properties such as the duration, type of events, the present casts, distance to the next events.

## 4 Experimental Results and Discussion

### 4.1 Parameter Setting

For experiments, we used 17 multimedia contents including *the Star Wars series: 6 episodes*, *the Lord of Rings series: 3 episodes*, and *the Harry Porter series: 8 episodes* in order to extract character network and analysis. Characters in these multimedia content were annotated according to time they appeared and disappeared during multimedia content playing time. The multimedia content had been indexed by analyzing character network using social network analysis technique with co-occurrence and total time that these appeared together during the multimedia content playing time and these were analyzed using character classification and social network evaluation strategies through the use of the score measurement. The system was implemented in Java using Vlcj API<sup>1</sup>; Gephi API<sup>2</sup> and Prefuse API<sup>3</sup>.

### 4.2 System Proposed and Evaluation Results

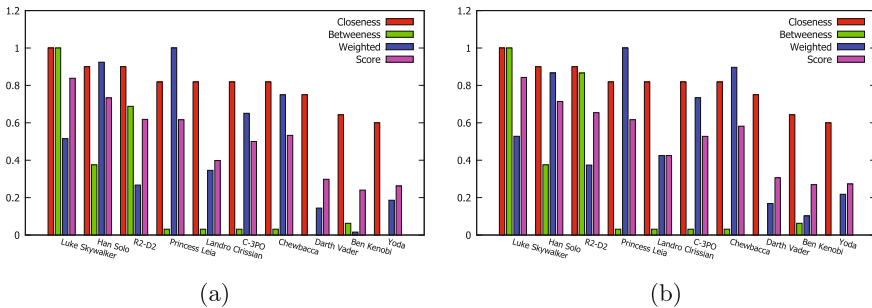
By applying the character's occurrences model, we implemented a system for annotating the appearance of characters during playback, in which we

<sup>1</sup> Caprica. <http://www.capricasoftware.co.uk/projects/vlcj/index.html>.

<sup>2</sup> Gephi API. <https://gephi.org/docs/api/>.

<sup>3</sup> Prefuse API. <http://prefuse.org/>.

assumed that the appearance of the characters in a multimedia content as the “onscreen” visually appearing. In this regards, each character’s appearance distribution will be calculated by the time which character appears (start time) and disappears (end time). Besides, we calculate the co-occurrence time distribution and total co-occurrence time of characters by using Definitions 4 and 5. Based on these results, we extracted a character network from the given multimedia content and applying measurement techniques to determine the protagonist and classifying characters into such class (by applying Definitions 9, 10, and 11). Results of these steps are described as the following (Fig. 2).



**Fig. 2.** The centrality of the characters in the multimedia content The Empire Strikes Back. (a) Total co-occurrence time; (b) Number of co-occurrence time

In order to index a multimedia content, we segment the sub-plots, extract character network using the annotations of characters in the multimedia content. This network were then analyzed by using Definitions 8, 10, and 11 to discover main character(s) and multimedia content’s main storytelling. Finally, indexing strategy were used for protagonist according to the character centrality measurement in order to detect sub-plots and main story line of the multimedia content that will then be used to produce the indexing version by using Definitions 13, 14, and 15.

Table 1 illustrates centralities of the characters in Star Wars Episode VI: Return of the Jedi. The results show that the main characters in the multimedia content have high degree of centrality and hold important roles in the multimedia content. The centralities of these results are combined by using Definition 9 to find that Luke Skywalker had the highest values. In the IMDB database ranking list about cast and crew of this episode, Luke Skywalker is main character who plays an important role to tell main story of this multimedia content. The centralities measurements are used to identify the main characters of the multimedia content in a list form, as be shown in Table 1. These characters are known as protagonists in the multimedia content. These results matched with those found in the IMDB’s database casts and crews ranking list.

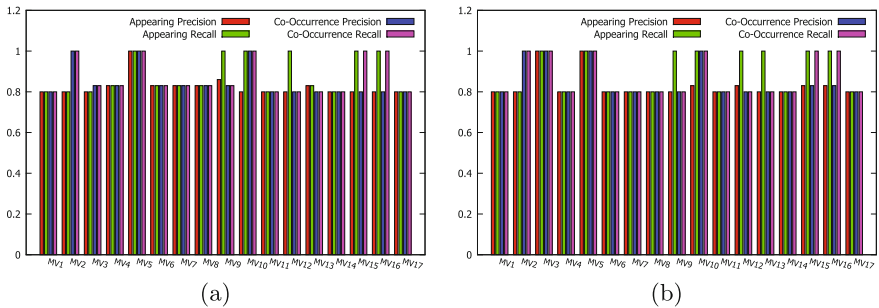
Figure 3 illustrate the results of characters classification, which are segmented characters into significant class and minor class. These results shown that our

**Table 1.** The centrality measure from the Star Wars Episode V. The Empire Strikes Back

Total co-occurrences time of characters					
ID	Character name	Closeness	Betweenness	Weighted	Score
1	Luke Skywalker	1.0000	1.0000	0.5153	0.8384
2	Han Solo	0.9000	0.3754	0.9243	0.7332
3	R2-D2	0.9000	0.6880	0.2673	0.6184
4	Princess Leia	0.8182	0.0311	1.0000	0.6164
5	Landro Clrissian	0.8182	0.0311	0.3459	0.3984
6	C-3PO	0.8182	0.0311	0.6510	0.5001
7	Chewbacca	0.8182	0.0311	0.7494	0.5329
8	Darth Vader	0.7500	0.0000	0.1442	0.2981
9	Ben Kenobi	0.6429	0.0628	0.0156	0.2404
10	Yoda	0.6000	0.0000	0.1860	0.2620

Number of co-occurrences time of characters					
ID	Character name	Closeness	Betweenness	Weighted	Score
1	Luke Skywalker	1.0000	1.0000	0.5281	0.8427
2	Han Solo	0.9000	0.3754	0.8663	0.7139
3	R2-D2	0.9000	0.8663	0.3742	0.6541
4	Princess Leia	0.8182	0.0311	1.0000	0.6164
5	Landro Clrissian	0.8182	0.0311	0.4247	0.4247
6	C-3PO	0.8182	0.0311	0.7337	0.5277
7	Chewbacca	0.8182	0.0311	0.8966	0.5820
8	Darth Vader	0.7500	0.0000	0.1685	0.3062
9	Ben Kenobi	0.6429	0.0628	0.1022	0.2693
10	Yoda	0.6000	0.0000	0.2180	0.2727



**Fig. 3.** Precision and recall of character classification. (a) Total time appearing together; (b) Number of co-occurrence

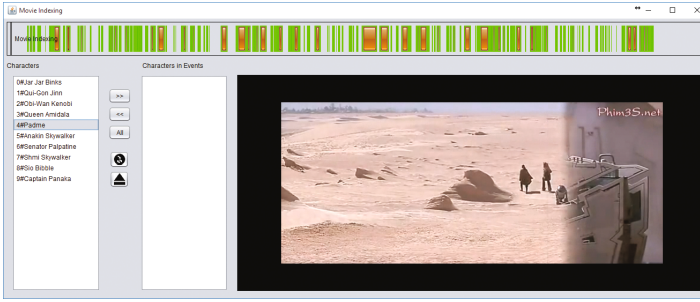


Fig. 4. Indexing module

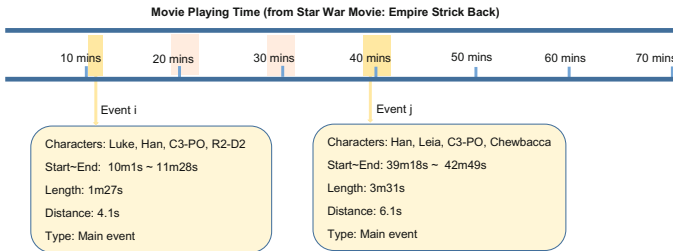


Fig. 5. An example of indexed multimedia content

proposed approach that using score of characters have a performance analysis. The main character class exhibited a precision of 83% and a recall of 89% on average. The minor character class exhibited a precision of 85% and a recall of 87%, on average for the total time co-occurrence analysis. In other analysis, the number of character co-occurrences are counted, and the main character class accomplished a precision of 82% and a recall of 88%, on average. The minor character class achieves a precision of 85% and a recall of 87%, on average. Compare to the proposed method in [9], our results showed the higher accuracy rate and performance. Besides, this method could not identified the protagonist of the given multimedia content.

Figure 4 illustrates multimedia content indexing system. Based on the character network, we could explore main storytelling and discover main character(s) of the multimedia content based on characters appearance characteristics. Using analyzing data from CoCharNet (as the characters appearance, relationships among characters and social network analyzing method) [6] and Definitions 13 and 14, we vents of the multimedia content into certain groups: Main events contain which belong to main story of the multimedia content, in which, include main character(s), sub events, in which, including minor characters and hybrid events, in which, contain both of main and minor characters. Compare to contents based method [4], this proposed approach is used audio and visual sources to extract high-level semantic cues as a set of events type (2-speaker dialogs,

multiple-speaker dialogs, and hybrid events). This proposed approach is based on speaker identification which had average of 88% classification accuracy only and it can not detect main characters and can not classification events based on the importance of them.

## 5 Conclusion

Story-based storytellings are central concerns of multimedia content theories. A storytellings are provided the imagination of the story while encouraging the audience. Current approaches are focused on discovering characters, their relationships based on content-based techniques, in which, using features object recognition and dialogs analysis. However, current approaches have some limitations in accuracy rate such as the performance and correction rate of features recognition and lacking of dialogs. In this paper, we proposed a new approach to index which are based on time intervals of characters, a new model for extracting temporal characteristics, in addition, we proposed a method to extract main storytellings which are based on character network and time temporal of characters. Results of analysis stages are used for indexing. Our experiments with 17 multimedia content and more than 2000 min have shown that this method can lead a better level of understanding and indexing extractions. However, this work has some limiation in the considering of characters' occurrence "onscreen" visually. Next period can be achieved by developing analysis performances based on applying other techniques in story-based.

## References

1. Bordwell, D.: *The Way Hollywood Tells It: Story and Style in Modern Movies*. University of California Press, Berkeley (2006)
2. Monaco, J.: *How to Read a Film: The Art, Technology, Language, History and Theory* Offilm and Media. Oxford University Press, Oxford (1977)
3. Del Fabro, M., Böszörményi, L.: State-of-the-art and future challenges in video scene detection: a survey. *Multimed. Syst.* **19**(5), 427–454 (2013). doi:[10.1007/s00530-013-0306-4](https://doi.org/10.1007/s00530-013-0306-4)
4. Li, J., Shrikanth, N., Kuo, C.C.J.: Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans. Circuits Syst. Video Technol.* **14**(8), 1073–1085 (2004). doi:[10.1109/TCSVT.2004.831968](https://doi.org/10.1109/TCSVT.2004.831968)
5. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. *J. Vis. Commun. Image Represent.* 121–143 (2008). doi:[10.1016/j.jvcir.2007.04.002](https://doi.org/10.1016/j.jvcir.2007.04.002). Academic Press, Inc., Orlando, FL, USA
6. Tran, Q.D., Hwang, D., Jung, J.J.: Movie summarization using characters network analysis. In: *Proceedings of 7th International Conference on Computational Collective Intelligence Technologies and Applications*, pp. 390–399. Springer (2015). doi:[10.1007/978-3-319-24069-5\\_37](https://doi.org/10.1007/978-3-319-24069-5_37)
7. Tran, Q.D., Hwang, D., Lee, O.J., Jason, J.J.: Exploiting character networks for movie summarization. *Multimed. Tools Appl.* 1–13, (2016). doi:[10.1007/s11042-016-3633-6](https://doi.org/10.1007/s11042-016-3633-6). Springer



8. Tran, Q.D., Hwang, D., Jason, J.J.: Character-based indexing and browsing with movie ontology. *J. Intell. Fuzzy Syst.* **32**(2), 1229–1240 (2017). doi:[10.3233/JIFS-169122](https://doi.org/10.3233/JIFS-169122)
9. Weng, C.-Y., Chu, W.-T., Wu, J.-L.: Rolenet: movie analysis from the perspective of social networks. *IEEE Trans. Multimed.* **11**(2), 256–271 (2009). doi:[10.1109/TMM.2008.2009684](https://doi.org/10.1109/TMM.2008.2009684)

# Plot-Creation Support System for Writing Novels

Atsushi Ashida<sup>1</sup>(✉) and Tomoko Kojiri<sup>2</sup>

<sup>1</sup> Graduate School of Science and Engineering, Kansai University, Suita, Japan  
k088944@kansai-u.ac.jp

<sup>2</sup> Faculty of Engineering Science, Kansai University, Suita, Japan  
kojiri@kansai-u.ac.jp

**Abstract.** When writing a novel, authors often create a framework called a plot. However, since a plot's format is not defined, they must create it by their own styles. If sufficient information for completing a novel is not included in the plot, especially by beginning authors, they might be unable to complete their novels. This study supports beginning authors who are creating plots for their novels. We analyze the items that must be considered in the process of writing a novel and propose a plot-construction model. In addition, we develop a system that support creating plots by introducing a story in the plot-construction model. A story corresponds to the events ordered along a time sequence of the narrative world.

**Keywords:** Novel writing · Plot · Creativity support · Plot-creation model

## 1 Introduction

Creativity is the ability to make new things and has recently received attention as one of the 21st century skills [1]. Writing novels is one activity that needs creativity. Authors need to construct a novel's framework and represent it by language in sentences. By writing novels, not only generating ideas but also writing ability is cultivated. However, some author especially beginners, cannot finish their novels, often because the contents to be written are not sufficiently organized in advance.

In writing novels, before writing sentences, some authors consider what contents to write about in what order and summarize them as a plot. A plot is a novel's framework that describes the contents that must be told, such as the novel's settings and the important events that occur in it. Since a plot's description format is not uniquely defined, therefore, how to describe the details depends on individual authors. Some authors fail to derive sufficient information for writing a novel, and so sometimes the plot's contents are not connected naturally or might even be contradictory. As a result, authors often abandon their novels. To cope with such problems, this study aims at supporting creating plots for writing novels. First, we define the necessary contents to consider for creating plots. Based on the results, we develop a system that supports plot-creation, especially for beginning authors.

---

This paper is submitted to the special session: 2.

Several researches have focused on generating sentences. Liu, et al. developed a system which generates sentences automatically by combining information from existing texts using a corpus and ontology [2]. Some studies focused on generating novels automatically by preparing knowledge for creating sentences [3, 4]. Automatic plots generation has also been studied [5]. These studies made computers that produce novels instead of helping human authors create novels. We think that the difficulty of writing novels is creating an original story and combining the information of existing novels does not lead to an original story. In order to write novels with original stories, we think that it is important to develop the ability of authors to create an original story. As idea-inducement support for novel writing, Kainuma, et al. developed a virtual space in which users can murmur ideas and allows others to view such comments [6]. This approach requires the cooperation of others and also fails to support the organization of ideas conceived by authors to form a plot. Watanabe, et al. analyzed the components of novels and developed a novel-writing support system by making authors describe them [7]. However, the components of novels do not equal what authors are considering while they are writing novels. For example, a novel's message, which is what an author wants to convey through her novel, sometimes does not appear explicitly in it. To support the creation of novels, we must address their contents and their items that are considered during the creation process and do not appear in the final, written version.

In this study, we propose a plot-construction model as a thinking scheme to write novels. Our plot-construction model expresses what authors are thinking about in creating a plot and its relationships. As a plot-creation support, we also develop a system in which authors can express a plot and a story in plot-construction model. Story represents the events that are ordered along the time sequences of the narrative world. By considering story in plot-creation, plot contradictions can be easily detected. Since the order of events in the plot are not always identical as those in the story and intriguing plots are sometimes created by changing the order of events from the narrative world, our system provides an environment that allows authors to think about the sequence of events in their plots while observing the story's events.

## 2 Process of Writing Novels

Figure 1 shows the novel-writing process. The first phase is getting an idea. When writing novels, authors generate key ideas, for example, specific phrases, characters, or themes. The second phase is planning, where authors construct a novel's structure so its key ideas coalesce into a novel. After such planning, a plot is created. The last phase is writing, where authors transform the plot's contents into language and sentences.

In the writing phase, if the plot's contents are insufficiently described, authors provide additional sentences to complete them. However, authors sometimes have difficulty writing because their sentences are inconsistent with the existing contents, and so they are not able to complete their novels. Therefore, for writing a novel, it is important to sufficiently prepare a plot in the planning phase. Since authors create plots by their own formats, they have difficulty noticing the lack of content.

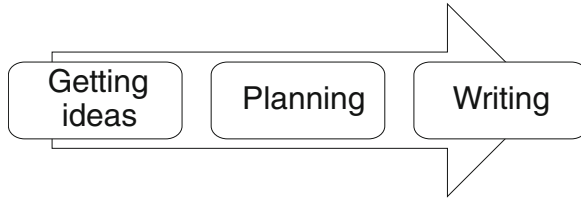


Fig. 1. Writing process of novels

### 3 Plot-Construction Model

To support plot-creation, we must clarify what authors should consider in the process of creating plots. We define the contents considered in the plot-creation process and their relationship as a plot-construction model (Fig. 2). In addition to the plot’s structure itself, the model consists of a message, a story, a strategy, and a reader model.

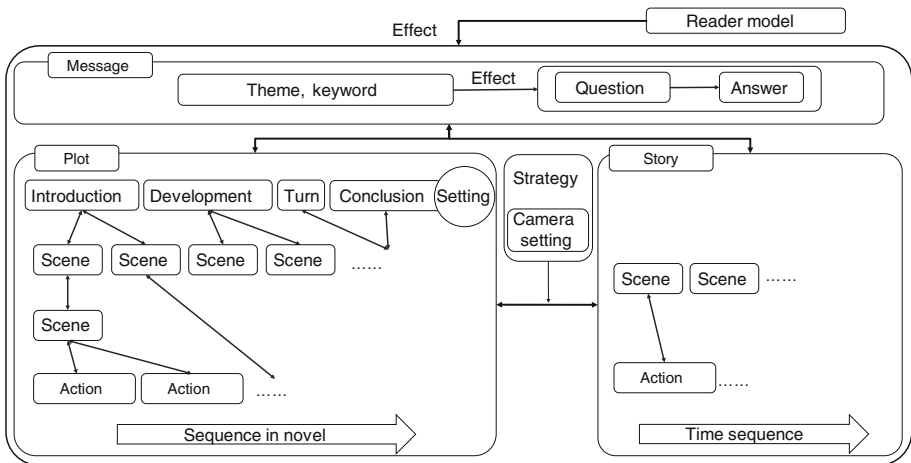
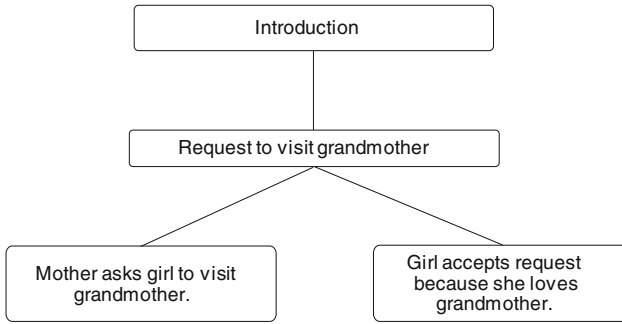


Fig. 2. Plot-construction model

Authors sometimes have a central question that they address and answer throughout their novel. Like this central question, authors sometimes have messages that they are interested in promulgating in their novels. Messages are sometimes derived from themes or keywords. Therefore, a message in the plot-construction model is composed of questions and answers; themes and keywords are triggers for deriving them.

Novels have target readers. A novel’s contents might change according to its readers. For instance, the using words or sentence structure undoubtedly differs for primary school students with low-reading ability and adults with high-reading ability. In our plot-construction model, target readers are expressed by a reader model.

Plot represents a novel’s framework, and creating a plot is the final goal of the plot-construction model. The plot holds scenes and their settings, which describe



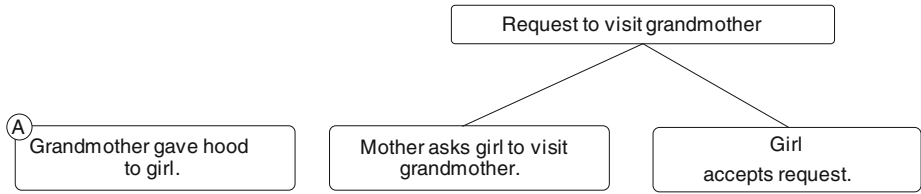
**Fig. 3.** Example of plot in Little Red Riding Hood

characters and the places where scenes occur. The plot holds scenes as a hierarchical structure. Based on this hierarchy, the contents become more specific. Since lower scenes show the details of upper scenes, the settings of upper scenes are inherited to lower scenes. In lower scenes, a setting unique to them is described. In the same hierarchy, the most left scene comes the first in the novel and the most right scene is the last one. The most detailed scenes are called actions. In the writing phase, actions are converted into sentences. In our plot-construction model, as a basic structure of novels, the following steps are adopted: introduction, development, turn, and conclusion. The scenes in the first layer correspond to introduction, development, turn, and conclusion scenes.

The order of the actions in a novel and in its narrative world is not necessarily identical. The differences in this order may be one interesting aspect of a novel. To show a novel's actions in an interesting order, authors need to grasp the actions in novels and their narrative worlds. In our plot-construction model, the actions that occur along a time sequence of the narrative world are called a story. Similar to plot, a story consists of actions and scenes that summarize such actions. Also, every actions and scenes have settings. However, no element expresses the basic structure of novels: introduction, development, turn, and conclusion. There is a correspondence between the actions in a plot and those in a story, but not all of the actions in a story need to exist in a plot.

When constructing a plot based on a story, actions, which are selected from the story and their orders, are decided by a strategy, which is the author's intention for how much information to give to readers. Camera setting is one element in a strategy that corresponds to the viewpoint from which a novel is written.

An example that represents a plot-construction model is shown using the Little Red Riding Hood, whose reader model is preschool children. One of its more obvious messages is that evil people are ultimately punished. In this message, the question might be, "What happens when someone does bad things to good children?" and an answer might be: "they lose." Figures 3 and 4 show the plot and the story of the tale's first part; the girl is sent by her mother to visit her grandmother. Since Little Red Riding Hood is described from a third-person perspective, the camera setting is a third-person's perspective. Depending on the strategy, the plot sequence is not as same as the



**Fig. 4.** Example of story in Little Red Riding Hood

story sequence. In Fig. 3, there is the action of “Girl accepts request because she loves grandmother.” There might be the scene in the story that represents why girl loves grandmother show in node A in Fig. 4. By adding node A to the story, a scene of girl’s recollection that grandmother gave hood to girl can be added to the plot. The setting describes the two main characters: a young girl and her mother. The grandmother’s house is far from the girl’s house.

## 4 Plot-Creation Support

We propose a support method for plot-creation based on a plot-construction model, which has four elements in addition to plot. Among them, the message, the reader model, and the strategy are generally considered in the first stage, and the story is addressed while creating the plot details. This study’s target is authors who are able to start thinking plots but cannot complete their novels. Such targets may face difficulties creating plot details. Therefore, this study supports plot-creation by convincing authors to consider story and to represent plot which based on plot format.

Some authors create plots with insufficient contents and without awareness of the story’s existence. Defining the plot and story format and providing an environment for representing them might support such authors by trial and error. For authors to be aware of insufficient contents or identify conflicting contents, not only writing down a plot and a story but also connecting their actions might be effective. Therefore, we developed a system for representing plot and story while associating their actions with each other.

Figure 5 shows the overall framework of our system, which consists of an interface in which the author can input a plot and a story and associate them. It also stores the plot and the story created by the author.

The interface provides an environment in which authors can input plot and story contents. It also visualizes the input contents so that authors can easily organize them. Since plot is represented as a hierarchical structure, it must also be expressed with a hierarchical structure in the interface. Plot consists of two types of nodes: actions and scenes. Both nodes have their respective settings. On the other hand, in the story, actions should be arranged along a time sequence. To create plot, authors must understand the actions in the story, not the scenes. Therefore, the interface provides an environment to input actions into the story and allows authors to relate actions in the plot and corresponding actions in the story. By organizing the information with such an interface, authors can create a plot by considering actions that occurred in the narrative world.

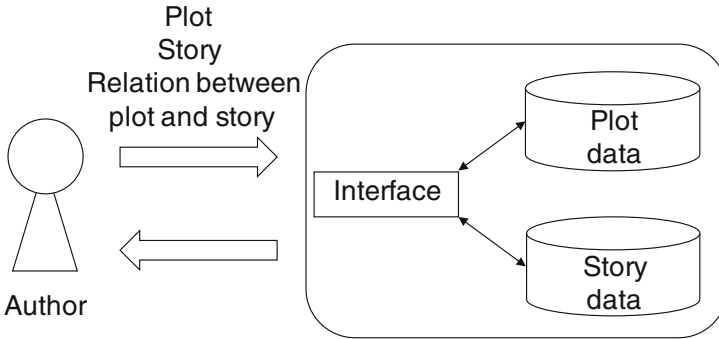


Fig. 5. Overall framework of plot-creation support system

## 5 Prototype System

We have developed a prototype system to support plot-creation. Representations of the story's time sequence and the plot's hierarchical structure are implemented in vis.js, and the other functions are implemented in JavaScript and jQuery. The system operates as a single page web application. The system process only works on the client side, not on the server side.

Figure 6 shows the system's interface, which consists of the following parts: plot, story, node editing, and a plot/story linking. The plot part creates the plot's hierarchical structure. The story part is used for arranging the time sequence of the actions in the narrative world. In the node editing part, the contents of each plot and story node can be input and edited. In addition, file operations as saving and reading, FAQ functions, and tutorials are prepared for each tab.

When the system is started, a root node and its child nodes, which corresponds to the novel's basic structure, including introduction, development, turn, and conclusion, are shown in the plot part. Authors can add a new node by clicking on the existing node to which the new node will be attached. The nodes of the basic structure of the novel, scenes, and actions are discriminated by colors. By dragging the nodes and moving them to the horizontal direction, their positions are changed within the same hierarchy. By dragging the background of the plot part, the focus of the plot tree in the interface can be changed. In addition, the plot tree can be zoomed out or in by a mouse-wheeling operation.

The contents of each node are input and edited on the "plot edit" tab of the node editing part. When a node is selected in the plot part, the details of the scene or the action and its setting are displayed on the plot edit tab. The setting of the parent node is inherited to the child node; when creating a new node, the setting of its parent node is set as its initial state. The selected plot nodes can also be removed with the "delete node" button in this tab. Authors cannot create, delete, move, or change the nodes of the novel's basic structure, such as introduction, development, turn, and conclusion.

In the story part, the actions in the story are expressed on a timeline that corresponds to the horizontal axis. The timeline represents the time flow from the left to the

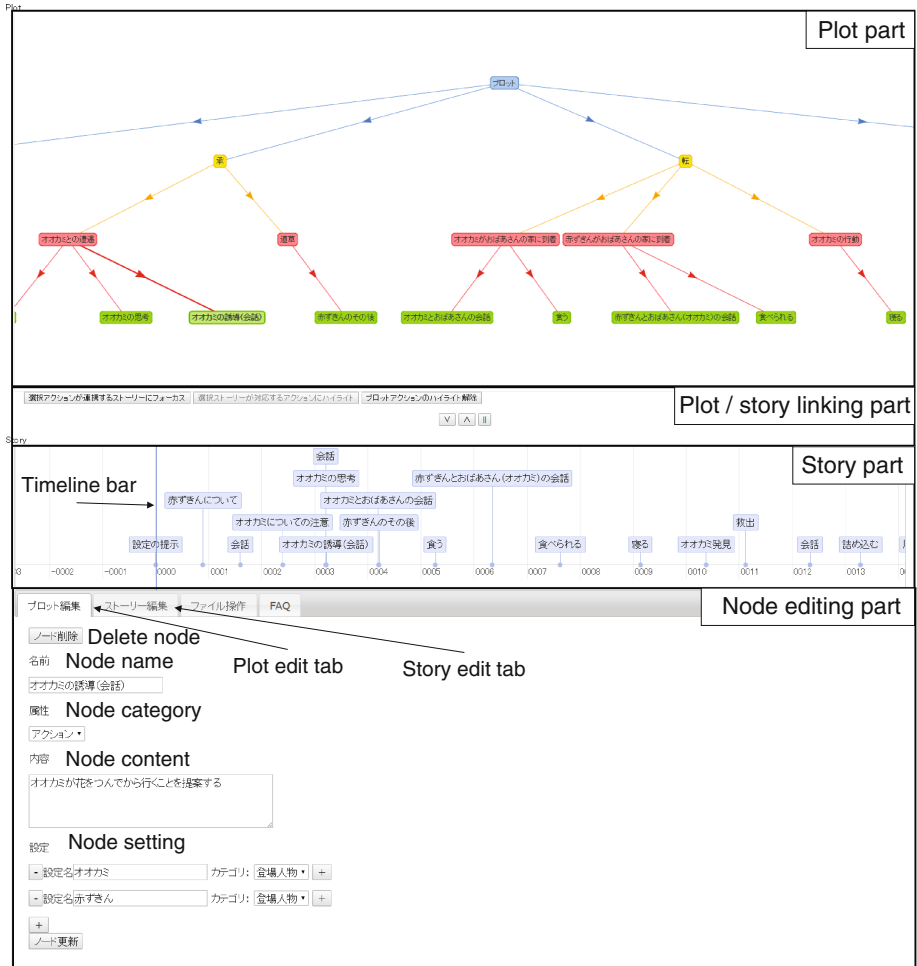


Fig. 6. System interface

right based on the numbers on the axis. Since grasping order of actions is important rather than occurrences time of every actions, the values on the time axis do not represent concrete year, day or time. They are just numbers and authors can use them with free interpretation. When the timeline is clicked, a new node is added to the clicked position. By clicking on the created story node, the author can input and edit its contents. Contents are displayed in the “story edit” tab of the node editing part. Authors can also delete the story node by clicking on the x button next to the node.

In the plot/story linking part, authors can correlate the nodes of the plot with the story of the same actions, insert the actions of the story into the plot, and insert the actions of the plot into the story. Since the story’s actions may appear more than once in the plot, the actions of the plot and those of the story have a many-to-one relationship. To make it easier for authors to grasp the relationship between the actions in



plots and stories, the corresponding actions of the plot or story are highlighted by selecting the action of the plot or the story and clicking on the button in the plot/story linking part.

## 6 Evaluation

### 6.1 Experimental Setting

We experimentally evaluated the effectiveness of our developed system for plot-creation. The objective of this experiment was to evaluate whether beginning authors could successfully create the complete plot using our system. Therefore, the usability of the system was mainly focused and the quality of the created plot was not evaluated. In addition, this experiment had not evaluated the effectiveness of our system for writing novels sentences.

In the experiment, we have asked nine undergraduate/graduate students who were interest in writing short story to create plots using our system. We first explained plot and story and how to use our system. Next, we asked participants to create plots. They were given one hour time limit to create a plot, but we allowed them additional time if they wanted.

After creating a plot, participants filled out questionnaires whose items are shown in Table 1. Items A and B asked the experience of participants regarding to the reading or writing novels. Items C and D asked the effectiveness of our system for creating a plot and item E is for the usability of the system.

**Table 1.** Questionnaire items

Items	Contents	Choices
A	Have you ever written a short story?	1. Yes, I have 2. I tried to write one, but I couldn't finish it 3. I thought about writing a short story 4. No
B	Describe your habits of reading books	1. I read daily 2. I read if I have time 3. I do not read much 4. I do not read at all
C	Were you able to smoothly create a plot? If not, explain why	1. Yes 2. No
D	Which part of the system was useful for plot-creation? (Multiple answers are allowed.)	1. Story part 2. Plot part 3. Plot/story linking part 4. Node editing part (setting)
E	Was the system easy to use?	1. Yes 2. No

**Table 2.** Questionnaire results

Items	Choices			
	1	2	3	4
A	0	3	2	4
B	4	3	1	1
C	6	3		
D	1	7	2	1
E	7	2		

## 6.2 Result

Table 2 shows the questionnaire results. The numbers are the amount of participants who selected each choice.

From item A, six participants had never written short stories and three had not completely a short story. From item B's results, many participants read books daily. These results suggest that our participants are beginners to writing novels, but they do read them.

Item C shows that six participants smoothly created plots using the system. Two of them who failed to create a plot answered that "The system did not give me new ideas." Another participant answered that "I got some ideas for writing a novel, but I could not connect them well." Our system does not support the creation of new ideas. Therefore, these three participants were not the target users that our system expects. However, based on these opinions, we need a method for supporting such participants to derive new ideas in the future.

From item D, seven participants replied that the plot part was helpful for creating plots. During the experiment, five participants did not use the story part, or even if they did use it, the orders of the actions of the plot and the story were the same. One participant said that "I understood the actions in the narrative world while making a plot without describing the story," and another said that "I did not have to use the story part because the sequence of the plot's actions is the same as those in the narrative world." On the other hand, four participants used the story part and changed the orders of the plot and story actions. All of the plots created by these participants were mysteries. We believe that the relationship between the story and plot actions is different based on the genres. For example, in a mystery, facts without their causes are presented to readers in the beginning, and causal actions that occurred before are revealed at later parts. Therefore, the effectiveness of the story part is changed based on genre.

From item E, seven participants gave good evaluations of the system's usability. Two participants complained that "Understanding how to use the system was difficult because it had too many buttons" and "Using the system was too complicated." We must improve the interface not to disturb authors of creating a plot. For instance, buttons should be rearranged more intuitively and with fewer eye movements.

## 7 Conclusion

This paper proposed a plot-construction model that represents the necessary factors to consider for creating the plot of the novel. We also developed a plot-creation support system in which a plot can be created by using format of the plot and the story in the plot-construction model. The experimental results clarified the effectiveness of the plot part for creating the plot. Moreover, the effectiveness of organizing actions in the narrative world was observed for specific genres, such as mysteries. Currently, we have only evaluated whether beginning authors could successfully create the plot using our system. The quality of the created plots needs to be evaluated. In addition, the effectiveness of the system for writing novel sentences should also be evaluated.

In the current system, plots are created by assigning subordinate nodes under superordinate nodes. This process corresponds to a top-down thinking process. However, some authors might create a plot from actions. Our system does not support such a bottom-up thinking process. To support this thinking process, our system must be improved to create a plot from action nodes as well.

The current system only introduced a plot and a story in the plot-construction model. We do not support the entire plot-creation process including messages, reader models, and strategies in plot-construction model. Messages and reader models are often considered in the early stage of creating plot. Therefore, we should introduce these parts into the system to encourage authors to derive ideas for their novels. In addition, we must verify a validity of plot-construction model.

**Acknowledgements.** The work was supported in part by JSPS KAKENHI Grant-in-Aid for challenging Exploratory Research (16K12563).

## References

1. Griffin, P.E., McGaw, B., Care, E.: *Assessment and Teaching of 21st Century Skills*. Springer, Heidelberg (2012)
2. Liu, C.L., Lee, C.H., Ding, B.Y.: Intelligent computer assisted blog writing system. *Expert Syst. Appl.* **39**(4), 4496–4504 (2012)
3. Akimoto, T., Ogata, T.: A consideration of the elements for narrative generation and a trial of integrated narrative generation system. In: 2011 7th International Conference on Natural Language Processing and Knowledge Engineering, pp. 369–377 (2011)
4. Sato, S.: A challenge to the third Hoshi Shinichi award. In: *The INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pp. 31–35 (2016)
5. Sakuma, T., Ogata, T.: Story generation support system used the story theory of propp. In: *JSAI2005*, 3D3-04 (2005). [in Japanese]
6. Kainuma, S., Miyashita, H., Nishimoto, K.: Analyses of creation processes of novels where others' whims are exploited to in-spire an author's imagination. *IPSJ SIG Technical report*, vol. 2006, no. 24, pp. 113–120 (2006). [in Japanese]
7. Watanabe, T., Arasawa, R.: Computer-supported novel composition based on externalization. *Procedia Comput. Sci.* **35**, 1662–1671 (2014)

# An Improved Algorithm for Mining Top-k Association Rules

Linh T.T. Nguyen<sup>1</sup>, Loan T.T. Nguyen<sup>2,3</sup>, and Bay Vo<sup>4</sup>(✉)

<sup>1</sup> Faculty of Information Technology, Dong An Polytechnic,  
Binh Duong, Vietnam

thuylinh@dongan.edu.vn

<sup>2</sup> Faculty of Information Technology, Nguyen Tat Thanh University,  
Ho Chi Minh City, Vietnam

nttloan@ntt.edu.vn, nthithuyloan@gmail.com

<sup>3</sup> Faculty of Mathematics, Informatics and Mechanics,  
University of Warsaw, Warsaw, Poland

<sup>4</sup> Faculty of Information Technology,

Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam

bayvodinh@gmail.com

**Abstract.** This paper proposes an improved algorithm of TopKRules algorithm which was proposed by Philippe et al. in 2012 to mine top-k association rules (ARs). To improve the performance of TopKRules, we develop two propositions to reduce search space and runtime in the mining process. Experimental results on standard databases show that our algorithm need less time than TopKRules algorithm to generate useful rules.

**Keywords:** Data mining · Association rule mining · Top-k association rules · Rule expansion

## 1 Introduction

Previous methods for mining ARs such as mining traditional ARs [1, 18], mining non-redundant ARs [23], mining most generalization ARs [17], mining ARs with interestingness measures [19] are based on minimum support ( $\varepsilon$ ) and minimum confidence ( $\gamma$ ) thresholds. These approaches generate a huge of ARs when user specifies much small values of  $\varepsilon$ . Otherwise, they generate less ARs if the input value is large. Moreover, it is difficult to determine a suitable  $\varepsilon$  value in order to have enough ARs that users expect [6]. In 2012, Fournier-Viger et al. proposed an algorithm for mining top-k ARs is called TopKRules, and the users enter the number of ARs ( $k$ ) they expects. The result of this algorithm helps to solve the calculation of finding useful ARs. The process of finding rules is including left expansion and right expansion of antecedent and consequent of original candidates. However, this algorithm needs much runtime because of considering many redundant rules.

In this paper, we improve TopKRules algorithm based on the property of confidence formula and conditions for rule expansion to eliminate unnecessary rules.

## 2 Basic Concepts

### Definition 1: Transaction Database [2]

Given a finite set of items  $I = \{i_1, i_2, \dots, i_n\}$ . A transaction database  $D$  is a set of transactions  $T = \{T_1, T_2, \dots, T_m\}$  where each transaction  $T_d$  ( $1 \leq d \leq m$ ) has an unique identifier, called *Tid*. Each transaction  $T_d$  is subsets of items in  $I$ . Table 1 demonstrates a transaction database with  $T = \{1, 2, 3, 4, 5, 6\}$  and  $I = \{1, 2, 3, 4, 5\}$ .

**Table 1.** Transaction database

Transactions	Items
1	{1, 2, 4, 5}
2	{2, 3, 5}
3	{1, 2, 4, 5}
4	{1, 2, 3, 5}
5	{1, 2, 3, 4, 5}
6	{2, 3, 4}

### Definition 2: Association Rule [1]

Let  $X, Y \subset I$  and  $X \cap Y = \emptyset$ , an association rule  $X \rightarrow Y$  indicates that the survival of itemset  $X$  leads to the survival of itemset  $Y$  in these transactions.

### Definition 3: Support [6]

The support of an itemset  $X$ , denoted as  $sup(X)$ , is the number of transactions that contain  $X$ .

The support of a rule  $X \rightarrow Y$ , denoted as  $sup(X \rightarrow Y)$ , is the ratio of the number of transactions contains  $\{X, Y\}$  and the number of transactions in the database.

$$sup(X \rightarrow Y) = \frac{sup(X \cup Y)}{|T|} \quad (1)$$

### Definition 4: Confidence [6]

Confidence of an association rule  $X \rightarrow Y$ , denoted as  $conf(X \rightarrow Y)$ , is the ratio of the number of transactions contains  $\{X, Y\}$  and the number of transactions contains  $X$ .

$$conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} \quad (2)$$

Sample ARs from database in Table 1 are listed in Table 2.

**Table 2.** Extracted ARs from database in Table 1

Rule	Support	Confidence
$\{1\} \rightarrow \{2\}$	0.67	1
$\{2\} \rightarrow \{5\}$	0.83	0.83
$\{1, 2\} \rightarrow \{5\}$	0.67	1
$\{5\} \rightarrow \{1, 2\}$	0.67	0.8
$\{3\} \rightarrow \{4\}$	0.33	0.5

**Definition 5: Size of a Rule [6]**

The size of a rule  $X \rightarrow Y$  is  $p * q$  if  $|X| = p$  and  $|Y| = q$ . A rule of size  $p * q$  is called larger than a rule of size  $r * s$  if  $p > r$  and  $q \geq s$ , or  $p \geq r$  and  $q > s$ .

**Definition 6: Frequent Association Rule [6]**

Let  $r : X \rightarrow Y$  be an association rule,  $r$  is considered to be frequent if its support is larger than or equal to  $\epsilon$ .

**Definition 7: Valid Association Rule [6]**

Let  $r : X \rightarrow Y$  be an association rule,  $r$  is valid if it's frequent and its confidence is larger than or equal to  $\gamma$ .

**Definition 8: Tidset of an Itemset [6]**

The tidset of an itemset  $X$  is denoted as  $tids(X)$ .

$$tids(X) = \{t | t \in T \wedge X \subseteq t\}.$$

For example,  $tids(\{1, 2\})$  in database given in Table 1 is  $\{t_1, t_3, t_4, t_5\}$ .

**Definition 9: Tidset of a Rule [6]**

The tidset of a rule  $X \rightarrow Y$  is denoted as  $tids(X \rightarrow Y)$  and is defined as  $tids(X \cup Y)$ . The support and confidence value of  $X \rightarrow Y$  can be defined by presence tids as follows:

$$sup(X \rightarrow Y) = \frac{|tids(X \cup Y)|}{|T|} \quad (3)$$

$$conf(X \rightarrow Y) = \frac{|tids(X \cup Y)|}{|tids(X)|} \quad (4)$$

**Definition 10: Left Expansion of a Rule [6]**

Rule expansion on the left is a process of adding an item  $i \in I$  into the antecedent of the rule  $X \rightarrow Y$  in order to form new rule  $X \cup \{i\} \rightarrow Y$  which has greater size.

**Definition 11: Right Expansion of a Rule [6]**

Rule expansion on the right is a process of adding an item  $i \in I$  into the consequent of the rule  $X \rightarrow Y$  in order to form new rule  $X \rightarrow Y \cup \{i\}$  which has greater size.

**Definition 12: Top-k Association Rules [6]**

Mining top-k ARs from transaction database [6] is a task of finding a set  $L$  including  $k$  ARs such that their confidences satisfy  $\gamma$  and their supports are highest, i.e.,  $L = \{r \in D \mid \text{conf}(r) \geq \gamma \wedge \neg \exists s \notin L \mid \text{conf}(s) \geq \gamma \wedge \text{sup}(s) > \text{sup}(r)\}$ .

**Property 1 [6]:**

Let  $i \in I$ . For rule  $r : X \rightarrow Y$  and  $r' : X \cup \{i\} \rightarrow Y$ ,  $\text{sup}(r) \geq \text{sup}(r')$ .

**Property 2 [6]:**

Let  $i \in I$ . For rule  $r : X \rightarrow Y$  and  $r' : X \rightarrow Y \cup \{i\}$ ,  $\text{sup}(r) \geq \text{sup}(r')$ .

From Property 1 and Property 2, the support of an expanded rule is not greater than that of original rule. This also means the expanding an infrequent rule will always produce an infrequent rule (Definition 6) because all the frequent rules can be found by recursively performing expansions on frequent rules of size  $1^*1$ .

**Property 3 [6]:**

Let  $i \in I$ , for rule  $r : X \rightarrow Y$  and  $r' : X \rightarrow Y \cup \{i\}$ ,  $\text{conf}(r) \geq \text{conf}(r')$ .

## 3 Related Works

### 3.1 Mining Frequent Itemsets

Agrawal et al. proposed Apriori algorithm [2], which mines itemsets by using itemsets with  $k$ -items to discover itemsets with  $(k + 1)$ -items. Firstly, it scans the database to find frequent single items ( $L_1$ ). Secondly,  $L_1$  is used to find  $L_2$ , which includes frequent itemsets having 2-items. Then,  $L_2$  is used to find  $L_3$ . Doing similar steps until no more frequent itemsets are found. In 2004, Han et al. proposed FP-Growth algorithm [7] to mine frequent itemsets based on FP-Tree structure without candidates generation. FP-Tree is an extended of prefix tree structure that represents the transaction database in a compact and complete way. The tree nodes include only frequent length-1 items, and are ordered in such a way that more frequently occurring nodes, more excellently opportunity of sharing nodes. Each transaction in the database is mapped to one path in the FP-Tree. In FP-Tree structure, it is required to scan database twice. The first time, it calculates supports of each single items and discards infrequent items. Then this algorithm sorts frequent items in decreasing order based on their support. The second time, it processes one transaction at a time to create the FP-tree. FP-Growth partitions FP-tree based on the prefix. It scans the path of FP-Tree recursively to find the frequent itemsets.

### 3.2 Mining Top-K/Top-Rank-k Frequent Itemsets

Many approaches for mining frequent (closed) itemsets have been proposed [1, 2, 7, 9, 11, 16]. However,  $\epsilon$  value affects to the result of frequent itemsets. If users specify a small value of  $\epsilon$ , there are possible performance issues in mining a huge number of frequent itemsets. Otherwise, there is less number of frequent itemsets with large value of  $\epsilon$ . There are some studies on top-k frequent itemset mining. Pietracaprina and

Vandin proposed TopKMiner algorithm [12] to mine top-k frequent closed itemsets. This algorithm uses a priority queue structure  $Q$  for speed up process of mining. Tzvetkov et al. proposed TSP algorithm [15] to mine top-k frequent closed sequential patterns with length greater than or equal to  $min\_l$ .

Besides, some studies for mining top-rank-k frequent (closed) itemsets have been carried out. Mining top-rank-k frequent itemsets is a task for finding first  $k$  groups with highest supports, itemsets, which have same support, have same rank (or group). FAE (Filtering and Extending) algorithm [3] is the early approach of mining top-rank-k frequent itemsets. This algorithm ignores unexpected itemsets and selects valid itemsets to extend frequent itemsets. Fang and Deng also proposed VTK (Vertical mining of top-rank-k frequent patterns) algorithm [5]. This algorithm is more efficient than FAE algorithm since it calculates support for frequent itemsets without scanning database. In 2014, Deng proposed NTK algorithm [4] for mining top-rank-k frequent itemsets using Node-Lists data structure. This algorithm is more efficient than both FAE and VTK. Huynh-Thi-Le et al. proposed iNTK algorithm [8] in order to improve performance of mining frequent itemsets. Saif-Ur-Rehman et al. proposed Top-K Miner algorithm [14] for mining top-rank-k frequent itemsets using CIS-tree (candidate-itemsets-search tree). Nguyen et al. proposed an efficient strategy for mining top-rank-k frequent closed itemsets [10] by modifying DCI-plus algorithm [13] and dynamic bit vectors technique [16].

### 3.3 Mining Top-k Association Rules

There are some studies related to top-k association rule mining, such as k-optimal rules discovery by Webb and Zhang [20], filtered top-k ARs discovery by Webb [21], Mining top-k fault tolerant ARs in data streams [22]. Recently, Fournier-Viger et al. proposed an approach of mining top-k ARs based on TopKRules algorithm [6] which extracting confidence ARs having highest support. TopKRules algorithm helps to resolve the problem of  $\varepsilon$  value determination in order to mine top-k rules which have highest support and satisfy user's specified  $\gamma$  value. However, this algorithm still has issue of runtime performance because of many candidates generation during rules expansion.

## 4 Proposed Algorithm

**Proposition 1:** If a rule has lexically greatest item in the antecedent equal to lexically greatest item in itemset, it cannot be expanded its antecedent. It is also true with right expansion. This is obvious to the candidate search method for expansion.

**Proposition 2:** If a rule has confidence less than  $\gamma$ , it cannot be expanded its consequence. Let  $r'$  is expanded from  $r$  and confidence of the rule  $r$  is smaller than minimum confidence. We can easily prove the confidence of the rule  $r'$  is less than the minimum confidence.

The proposed algorithm is presented in Fig. 1. Initially,  $\varepsilon$  value is assigned to 0 (line 1). Then, the algorithm scans database  $T$  and saves tidset of each items, set of these single items are sorted in a semantic meaning, such as order by alphabet (line 2). This algorithm has two main steps:



- Step 1: Generate rules with size  $1*1$  by considering each pair of items  $i, j \in I$  (the set of distinct items form transactions) which satisfy  $\varepsilon$ . For each pairs  $(i, j)$ , it constructs two rules  $\{i\} \rightarrow \{j\}$ . and  $\{j\} \rightarrow \{i\}$  and adds the valid rule into a list  $L$  which contains current top-k rules (lines 4–6). For each pair rule  $\{i\} \rightarrow \{j\}$  and  $\{j\} \rightarrow \{i\}$ , if the antecedent contains greatest item (in semantic meaning), this rule can be only expanded on the consequence (cf. Proposition 1). Otherwise, it can be expanded on both antecedent and consequence, two rules are added into  $R$  (set of candidates which has capable of expansion) (lines 7, 8).
- Step 2: Afer generating rules having size  $1*1$ , it continues to expand to generate new rules from set of candidates  $R$ . Expansion process will be ended when there are no remaining candidates in  $R$  (line 9). It selects the rule has highest support and satifies  $\varepsilon$  in  $R$  then expand on the antecedent or consequence by calling procedure EXPANDL a EXPANDR (lines 12, 13). Then, it removes the considered rule out of set of candidates  $R$  (line 14) and removes rules having support less than  $\varepsilon$  (line 15).

---

**iTopKRules** ( $\gamma, k, T$ )

1.  $\varepsilon := 0; R := \emptyset; L := \emptyset$
  2. Scan database  $T$  and store tidset of each single items
  3. For each pair of items  $i, j \in I$  frequent, construct 2 rules:  $r_1 : \{i\} \rightarrow \{j\}$  and  $r_2 : \{j\} \rightarrow \{i\}$
  4. If  $sup(r_1) \geq \varepsilon$  then
  5.     If  $conf(r_1) \geq \gamma$  then AddTopK( $r_1, L, k, \varepsilon$ )
  6.     If  $conf(r_2) \geq \gamma$  then AddTopK( $r_2, L, k, \varepsilon$ )
  7. For each pair rules  $r_1$  and  $r_2$ , if the lexically of the largest item in the antecedent is equal to lexically greatest item in itemset  $\mathcal{I}$ , the expansion flag is turned off (only to be right expansion). Otherwise, expansion flag is turned on (can expand on both sides)
  8.      $R := R \cup \{r_1, r_2\}$
  9. While  $R \neq \emptyset$  do
  10. Choose the rule  $r \in R$  which has highest support.
  11. If  $sup(r) \geq \varepsilon$  then
  12.     If  $r.ExLR = true$  then EXPANDLR( $r, L, R, k, \varepsilon, \gamma$ )
  13.     Else EXPANDR( $r, L, R, k, \varepsilon, \gamma$ )
  14.  $R := R \setminus \{r\}$
  15.  $R := R \setminus \{\forall s \in R \mid sup(s) < \varepsilon\}$
- 

**Fig. 1.** iTopKRules algorithm

**The AddTopK Procedure (Fig. 2):** Initially, add valid rule  $r$  into  $L$ . If the number of rules in  $L$  is greater than  $k$ , the algorithm removes the rules having least support until  $L$  has  $k$  rules and updates  $\varepsilon$  value as the minimum support of rules in  $L$ .

---

**AddTopK**( $rule, L, k, \varepsilon$ )

1.  $L := L \cup \{rule\}$
2. If  $|L| > k$  then
3.   While  $|L| > k$  do
4.     Remove the rules having least support.
5.   Update  $\varepsilon$  as the minimum support of rules in  $L$ .

---

**Fig. 2.** AddTopK procedure

**EXPANDL and EXPANDR Procedures:** Input parameters for these procedures are rule having capable of expansion  $r, L, R, k, \varepsilon$  and  $\gamma$ . The expansion step of  $r$  is performed by adding single item  $p$  sequentially into the expanded side. Candidate items can be added to expand rule  $r$  is the itemsets which appear in same transactions with items in rule  $r$ , have lexically greater than that of all items in the expanded side and not appear in the remaining side.

The left expansion procedure **EXPANDL** is presented in Fig. 3. Sequentially, this algorithm adds item from candidate items into antecedent of rule  $r$  to form new rule  $r'$  which has greater size than that of  $r$ . If  $r'$  has support greater than or equal to  $\varepsilon$  and its confidence satisfies  $\gamma$  then adds  $r'$  into  $L$ . If  $r'$  has lexically greatest item in the

---

**EXPANDL**( $r, L, R, k, \varepsilon, \gamma$ )

1. Find all possible items to add into the antecedent. These items are the ones appear in same transactions with items in  $r$ , have lexically greater than all items in the antecedent, and do not present in the consequence.
2. With each item  $p$  from candidate items, try to add it into the antecedent of  $r$  to form new rule  $r'$ .
3.   If  $sup(r') \geq \varepsilon$  then
4.     If  $conf(r') \geq \gamma$  then AddTopK( $r', L, k, \varepsilon$ )
5.     If the lexically greatest item in the antecedent is the greatest item in itemset  $I$  then
6.        $r'.ExLR = false$
7.     Else  $r'.ExLR = true$
8.    $R := R \cup \{r'\}$

---

**Fig. 3.** EXPANDL procedure

antecedent which is greatest item in itemset, the rule only has capable of right expansion (cf. Proposition 1). Otherwise, this rule has capable of both sides expansion. Add this rule  $r'$  into  $R$ .

The procedure for expanding the consequence of a rule, EXPANDR (Fig. 4), scans and adds each item from candidate items sequentially into the consequence of rule  $r$  to form a new rule  $r'$ . The set of items to be considered contains all the items which appear in the same transactions with all items in rule  $r$ , have lexically greater than the greatest in the consequence of rule, do not present in the antecedent. If this new rule  $r'$  satisfies  $\varepsilon$  and  $\gamma$ , it is added into  $L$ . If the lexically greatest item in rule  $r'$  is the greatest item in itemset  $I$ , this new rule does not have capability of right expansion (cf. Proposition 1). Otherwise, it has capability of right expansion. If  $r'$  only satisfies  $\varepsilon$  and does not satisfy  $\gamma$ , this rule does not have any capability of expansion (cf. Proposition 2). Thus, it is not added into  $R$ .

---

**EXPANDR** ( $r, L, R, k, \varepsilon, \gamma$ )

1. Find all possible items to add into consequence. These items are the ones appear in same transactions with items in  $r$ , have lexically greater than all items in the consequence, and do not appear in the antecedent.
2. With each item  $p$  from candidate items, try to add it into the consequence of  $r$  to form new rule  $r'$ .
3. If  $sup(r') \geq \varepsilon$  then
4. If  $conf(r') \geq \gamma$  then
5. AddTopK ( $r', L, k, \varepsilon$ )
6. If the lexically greatest item in the consequence is not the greatest item in itemset then
7.  $r'.ExLR = false$
8.  $R := R \cup \{r'\}$

---

Fig. 4. EXPANDR procedure.

Proposed algorithm applied two proposition to eliminate unnecessary rules. Besides, the set  $R$  and  $L$  are sorted by support. Thus, runtime of *iTopKRules* is less than original algorithm. This is evident in the results of experiments on standard databases.

## 5 Experimental Results

### 5.1 Experimental Databases and Environments

The experiments were implemented and tested on a system with the following configuration: Intel Core I5-6200U 2.3 GHz (4 CPUs), 8 GB of RAM, and running Windows 10, 64 bit version. Our source code was implemented in Java with above

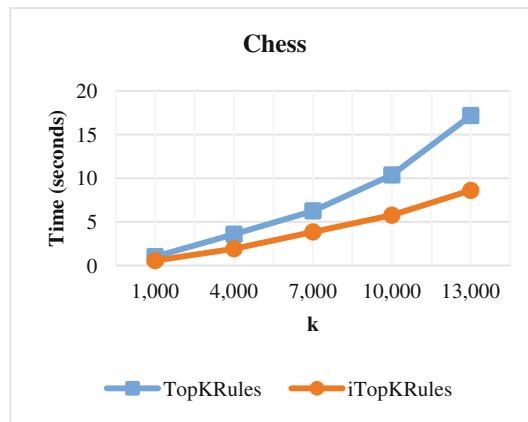
configuration. The standard databases used for testing were downloaded from <http://www.philippe-fournier-viger.com/spmf/>. The features of these databases are presented in Table 3.

**Table 3.** Testing databases

Name	No. of transactions	No. of items
Chess	3,196	75
Connect	67,557	129
Mushroom	8,416	128
Pumsb	49,046	7,116

## 5.2 Experimental Results

We executed our proposed algorithm with 4 standard databases presented in Table 3. We compared on time execution between TopKRules algorithm [6] and the proposed algorithm. The results are presented in Figs. 5, 6, 7 and 8. In the experiments, we executed both algorithm with fixed  $\gamma = 0.8$  and various value of  $k$  from 1,000 to 13,000.



**Fig. 5.** Runtime comparison between TopKRules and iTopKRules on Chess database.

Figures 5, 6, 7 and 8 indicate that iTopKRules is more efficient on runtime than that of TopKRules. The differences on runtime were not much when using small value of  $k$ , otherwise, when we mined top- $k$  rules using large value of  $k$ , there were much differences on runtime between TopKRules and iTopKRules algorithms.

It can be observed that the larger value of  $k$  we used, the faster iTopKRules had. In Fig. 5, with  $k = 1,000$ , both algorithms had nearly same execution time. However, with  $k = 4,000, 7,000, 10,000$ , iTopKRules needed less runtime than TopKRules needed.

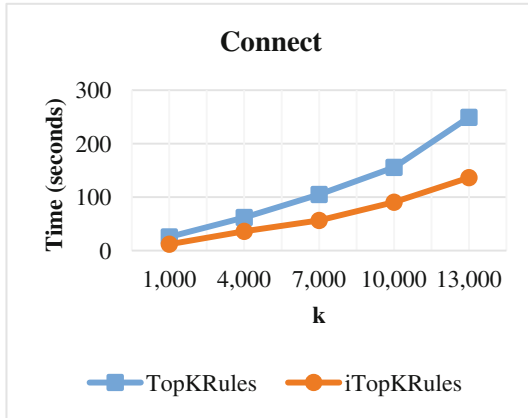


Fig. 6. Runtime comparison between TopKRules and iTopKRules on Connect database.

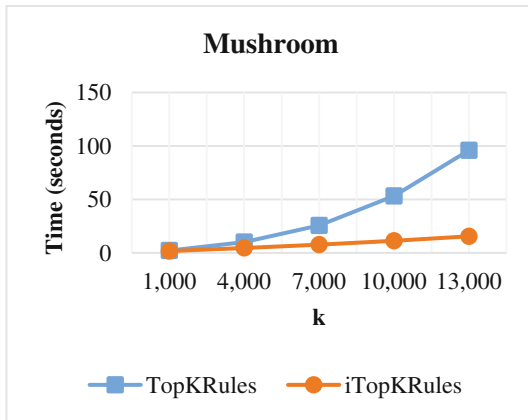


Fig. 7. Runtime comparison between TopKRules and iTopKRules on Mushroom database.

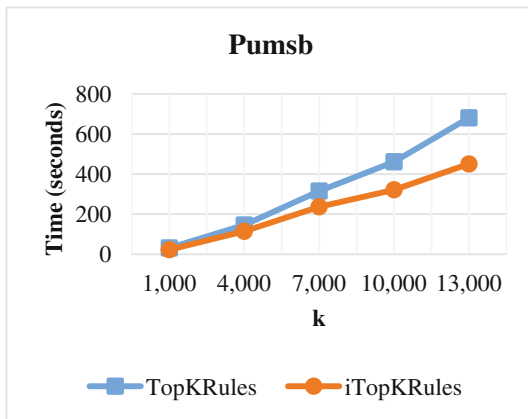


Fig. 8. Runtime comparison between TopKRules and iTopKRules on Pumsb database.

Especially, with  $k = 13,000$ , iTopKRules only took one half of runtime versus that of TopKRules. In Fig. 7, with  $k = 13,000$ , mining time on Mushroom database of iTopKRules was less five times than that of TopKRules.

## 6 Conclusions and Future Works

In this paper, we have proposed an efficient algorithm to extract top-k ARs. This algorithm solved the problem of difficulty to select suitable  $\varepsilon$  in order to have enough useful ARs. By applying property that only expand rules for rules having high confidence, our proposed algorithm improved runtime, especially with large expected number of ARs.

Currently, our work still needs  $\gamma$  value for mining. In the future work, we are going to investigate on mining top-k ARs without using  $\gamma$ , the rules in result will have highest confidence. Moreover, we will also investigate to mine top-k non-redundant ARs.

**Acknowledgments.** This work was carried out during the tenure of an ERCIM ‘Alain Bensoussan’ Fellowship Programme.

## References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings ACM International Conference on Management of Data, pp. 207–216. ACM Press (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
3. Deng, Z., Fang, G.: Mining top-rank-k frequent patterns. In: ICMLC 2007, pp. 851–856 (2007)
4. Deng, Z.H.: Fast mining top-rank-k frequent patterns by using node-lists. *Expert Syst. Appl.* **41**(4), 1763–1768 (2014)
5. Fang, G., Deng, Z.H.: VTK: vertical mining of top-rank-k frequent patterns. In: FSKD 2008, pp. 620–624 (2008)
6. Fournier-Viger, P., Wu, C.-W., Tseng, V.S.: Mining top-k association rules. In: Proceedings of the 25th Canadian Conference on Artificial Intelligence (AI 2012). LNAI, vol. 7310, pp. 61–73. Springer, Heidelberg (2012)
7. Han, J., Pei, H., Yin, Y.: Mining frequent patterns without candidate generation. *Data Min. Knowl. Disc.* **8**(1), 53–87 (2004)
8. Huynh-Thi-Le, Q., Le, T., Vo, B., Le, B.: An efficient and effective algorithm for mining top-rank-k frequent patterns. *Expert Syst. Appl.* **42**(1), 156–164 (2015)
9. Le, T., Vo, B.: An N-list-based algorithm for mining frequent closed patterns. *Expert Syst. Appl.* **42**(19), 6648–6657 (2015)
10. Nguyen, L.T.T., Trinh, T., Nguyen, N.T., Vo, B.: A method for mining top-rank-k frequent closed itemsets. *J. Intell. Fuzzy Syst.* **32**(2), 1297–1305 (2017)
11. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. *Inf. Syst.* **24**(1), 25–46 (1999)

12. Pietracaprina, A., Vandin, F.: Efficient incremental mining of top-k frequent closed itemsets. In: Tenth International Conference Discovery Science, pp. 275–280. Springer, Heidelberg (2004)
13. Sahoo, J., Das, A.K., Goswami, A.: An effective association rule mining scheme using a new generic basis. *Knowl. Inf. Syst.* **43**(1), 127–156 (2015)
14. Saif-Ur-Rehman, J., Ashraf, J., Habib, A., Salam, A.: Top-k miner: top-k identical frequent itemsets discovery without user support threshold. *Knowl. Inf. Syst.* **48**(3), 741–762 (2016)
15. Tzvetkov, P., Yan, X., Han, J.: TSP: mining top-k closed sequential patterns. *Knowl. Inf. Syst.* **7**(4), 438–457 (2005)
16. Vo, B., Hong, T.P., Le, B.: DBV-miner: a dynamic bit-vector approach for fast mining frequent closed itemsets. *Expert Syst. Appl.* **39**(8), 7196–7206 (2012)
17. Vo, B., Hong, T.P., Le, B.: A lattice-based approach for mining most generalization association rules. *Knowl.-Based Syst.* **45**, 20–30 (2013)
18. Vo, B., Le, B.: Mining traditional association rules using frequent itemsets lattice. In: International Conference on Computers and Industrial Engineering, pp. 1401–1406. IEEE Press (2009)
19. Vo, B., Le, B.: Interestingness measures for association rules: combination between lattice and hash tables. *Expert Syst. Appl.* **38**(9), 11630–11640 (2011)
20. Webb, G.I., Zhang, S.: K-optimal rule discovery. *Data Min. Knowl. Disc.* **10**(1), 39–79 (2005)
21. Webb, G.I.: Filtered top-k association discovery. *WIREs Data Mining Knowl. Discov.* **1**(3), 183–192 (2011)
22. You, Y., Zhang, J., Yang, Z., Liu, G.: Mining top-k fault tolerant association rules by redundant pattern disambiguation in data streams. In: International Conference Intelligent Computing and Cognitive Informatics, pp. 470–473. IEEE Press (2010)
23. Zaki, M.J.: Mining non-redundant association rules. *Data Min. Knowl. Disc.* **9**(3), 223–248 (2004)

# A Deep Architecture for Sentiment Analysis of News Articles

Dinh Nguyen<sup>1(✉)</sup>, Khuong Vo<sup>2</sup>, Dang Pham<sup>2</sup>, Mao Nguyen<sup>2</sup>,  
and Tho Quan<sup>1</sup>

<sup>1</sup> Ho Chi Minh City University of Technology, 268 Ly Thuong Kiet Street,  
District 10, Ho Chi Minh City, Vietnam  
{7140823, qttho}@hcmut.edu.vn

<sup>2</sup> YouNet Company, 2nd Floor, Lu Gia Plaza, 70 Lu Gia Street, District 11,  
Ho Chi Minh City, Vietnam  
{khuongva, dangpnh, maonx}@younetco.com

**Abstract.** In this paper, we present a deep architecture to perform aspect-level sentiment analysis for news articles. We combine some neural networks models proposed in various deep learning approaches, aiming at tackling specific issues commonly occurring for news articles. In this paper, we explain why our architecture can handle typically-long and content-specific news articles, which often cause overfitting when trained with neural networks. Moreover, the proposed architecture can also effectively process the case when the subject to be analyzed sentimentally is not the main topic of the concerned article, which is also a common issue when performing aspect-level sentiment processing. Experimental results with real dataset demonstrated advantages of our approach as compared to the existing approaches.

**Keywords:** Aspect-level sentiment analysis · News sentiment analysis · Deep learning · Convolution neural network · LSTM network · Word embedding network

## 1 Introduction

*Sentiment analysis* [1] or *opinion mining* [2] is the task that aims to infer the *sentiment orientation* in a document [3]. There are three levels of sentiment: (i) document-based level; (ii) sentence-based level; and (iii) aspect-based level. In document-based and sentence-based sentiment analysis, it is implicitly assumed that the analyzed document or sentence only discusses a single object. Recently, Hu and Liu (2004) proposed to perform sentiment analysis at *aspect-level* [4]. In this direction, apart from rating positive/negative sense of a text, the objects targeted by the mention (it may be a brand, a product or a feature) must also be identified.

Various works have been reported to handle this problem. *Topic modeling* was commonly applied in this case [5]. Mei et al. [6] proposed using *Probabilistic Latent Semantic Analysis (PLSA)*, while most recent works were based on *Latent Dirichlet Allocation (LDA)* [7–10]. Insight analysis [12] improves the performance by inferring *features* of the analyzed aspect.



Recently, with the emerging of *Treebank*, especially the famous Stanford Sentiment Treebank [12], sentiment analysis techniques using *deep learning* are increasingly adopted. Back to 1989, LeCun did propose an architecture of shared weights in neural networks [13]. Subsequently, suggested by Hinton [14], the idea of feeding a neural network with inputs through multiple-way interactions, parameterized by a tensor have been proposed for relation classification [15]. In the famous Stanford NLP tool [16], which was used commonly in the NLP research community, *RTNN* techniques have been applied on Stanford Sentiment Treebank and obtained remarkable performance [17]. Convolution-based techniques continue to be developed for sentiment analysis at sentence level [18]. To capture the relationship made by the appearance order of features, the *Recurrent Neural Network* (RNN) system such as *Long Short-Term Memory* (LSTM) has been used in combination with convolution processing to sentiment analysis for short text [19].

In this article, we concern sentiment analysis for the news articles at *aspect level*. That is, we want to determine the sentiment orientation for a specific object in the article, even though this object is not the main topic of the article, but only partially mentioned. The challenges of this task are as follows.

<p><b>Nhà mạng Mobifone khắc phục sóng yếu theo kiểu nửa vời? Đứt cáp quang AAG: Mạng internet còn chập chờn đến ngày 21/8</b></p> <p>Sau gần 2 tháng kể từ khi Báo Người tiêu dùng phản ánh về tình trạng sóng điện thoại yếu tại các khu chung cư cao tầng như HH1.2,3,4ABC, VP5.6 Tây Nam Linh Đàm (Hoàng Liệt, Hoàng Mai), CT12ABC KVKL (Đại Kim).</p> <p>...</p> <p>Người dân hết kiên nhẫn muốn chuyển mạng khác! Ngay sau khi nhà mạng <b>Mobifone</b> có thông tin phản hồi về tình trạng sóng kém chất lượng ở KĐT Linh Đàm và KĐT KVKL, phóng viên Báo Người tiêu dùng đã liên hệ để kiểm tra lại thông tin tại những nơi được cho là sóng Mobifone đều khỏe, thì thật bất ngờ <b>người dân lại phản hồi ngược lại</b></p> <p>...</p>	<p><b>Mobifone operators overcome weak signal problem in half-hearted style? AAG fiber disrupted: The Internet was intermittent until 21/8</b></p> <p>After nearly two months since the Consumer Report reflects on the state of weak electromagnetic waves in the area such as HH1.2,3,4ABC condominiums, Southwest VP5,6 Linh Dam (Hoang Liet, Hoang Mai) , CT12ABC KVKL (Dai Kim)</p> <p>...</p> <p>Impatient people want to move another network! Immediately after <b>Mobifone</b> networks have feedback about the status of poor quality waves at KDT Linh Dam urban area and KVKL, reporters contacted consumers to check the information in places where supposedly Mobifone waves are well, then <b>surprisingly people again reported the opposite opinion</b>.</p> <p>...</p>
---	---

Fig. 1. An article about optical fiber and Mobifone.

- The length of a news article is generally long. Thus, when trained by neural networks, it causes long execution process and extensive resource consuming.
- When information is published as a news article, it is quite specific to a certain topic, which would be discussed several times throughout the whole article. Thus, when trained by neural network, it is likely to cause the over-fitting phenomenon. For example, in Fig. 1 is an article about the optical fiber problem in Vietnam. This problem can negatively affect services provided by *Mobifone*, which is the subject supposed to be performed sentiment analysis. However, since this article mainly discusses about fiber problem, the trained neural network likely tends to “conclude” that all of issues related to fiber optic cable will cause a negative impact to *Mobifone*, which is not always the case.

- In some other cases, the subject to be analyzed is not necessarily the main subject of the article. For example, in Fig. 2 is an article about the singer Noo Phuoc Thinh with various negative terms (in red color). However, this article has a positive impact on *Mobifone* when this singer suggested users to use this service. If trained in a typical way, a neural network may tend to rate *neutral* or *negative* for all subjects mentioned in the article.

<p><b>Noo Phước Thịnh livestream tâm sự về 'The Voice'</b></p> <p>Ngày 11/2, t.rước khi "The Voice 2017" lên sóng, Noo Phước Thịnh đã livestream tâm sự về chương trình và team Noo cũng như một số thông tin hữu ích cho người hâm mộ.</p> <p>...</p> <p>Trước khi lên sóng, có nhiều ý kiến trái chiều cho rằng ngoài Thu Minh, 3 nghệ sĩ còn lại có <b>tuổi đời, tuổi nghề quá trẻ, chưa xứng đáng</b> đồng đội "ghê nóng" của một chương trình lớn như The Voice 2017.</p> <p>Dù đã khẳng định khả năng làm HLV của mình tại The Voice Kids 2016, Noo Phước Thịnh vẫn không tránh khỏi những lời <b>nhận xét khó nghe</b> từ dư luận. Tuy nhiên, nam ca sĩ tỏ ra rất tự tin và hào hứng với vị trí mới của mình tại The Voice 2017.</p> <p>...</p> <p>Bên cạnh đó, trong sự nghiệp của mình, Noo Phước Thịnh đã <b>hiều lần vấp ngã</b>, nên nam ca sĩ mong The Voice sẽ là nơi để truyền tải và giúp đỡ các bạn thí sinh bằng chính những vấp ngã mà anh đã trải qua.</p> <p>...</p> <p><b>Cẩn thận hơn, Noo Phước Thịnh còn hướng dẫn fan cách đăng ký sử dụng gói 3G giá rẻ của Mobifone, cụ thể: soạn DK FBI gửi 999 (3.000 đồng/ngày) để lướt Facebook, miễn phí data; soạn YT1 gửi 999 (10.000 đồng/ngày) để truy cập Youtube và FPT Play không giới hạn 3G.</b></p> <p>...</p>	<p><b>Noo Phuoc Thinh livestream talk about 'The Voice'</b></p> <p>11/2 days, before "The Voice 2017", Noo Phuoc Thinh has a livestream talk about the program and Noo team as well as some useful information for fans.</p> <p>...</p> <p>Before airing, there are mixed opinions that outside Thu Minh, 3 remaining artists is <b>too young, in experienced, not worthy</b> to sit "hot seat" of a big show like The Voice in 2017.</p> <p>Despite confirming his ability to coach at The Voice Kids 2016, Noo Phuoc Thinh still inevitable suffered from <b>harsh comments</b> from the public. However, the singer was very confident and excited about his new position at The Voice in 2017.</p> <p>...</p> <p>Besides, during his career, Noo Phuoc Thinh has <b>repeatedly stumbled</b>, so the singer wishes The Voice will be a place for the transmission and help the contestants with the same situations.</p> <p>...</p> <p><b>More carefully, Noo Phuoc Thinh longer guided subscribing fans use cheap 3G package Mobifone, namely: preparation FBI send DK 999 (3,000 VND / day) to surf Facebook, free data; YT1 compose posts 999 (10,000 VND / day) to access Youtube and FPT Play unlimited 3G.</b></p> <p>...</p>
---	--

**Fig. 2.** - An article mentions Mobifone with positive sentiment, but this service is not the main subject of the article.

To tackle these problems, we propose an architecture performing deep sentiment analysis for news articles. In this architect, we combine various methods which have been proposed in the relevant research, to address the above issues as follows.

- We use the Word Embedding model to reduce the word dimensionality before proceeding to the subsequent neural networks. This is to reduce resource consumed by the learning system.
- We use the LSTM approach to handle the order relationships between the sentences in articles. In particular, this model will “forget” the current temporary learning results if finding that this current result is deemed irrelevant to the analyzed object. Thus, for the article in Fig. 2, the negative information about Noo Phuoc Thinh will be “forgot” when the system starts processing to the discussion of Mobifone.

- Finally, we use multiple layers of *drop-outs* in the whole architecture. This is based on a proposal from [20], where drop-out factors can be used to prevent over-fitting on any back propagation learning system.

The rest of this paper is organized as follows. In Sect. 2 we present the background of CNN, Word Embedding, and LSTM. Section 3 presents our proposed deep architecture. Section 4 discusses experimental results. Finally, Sect. 5 concludes the paper.

## 2 Background

### 2.1 Convolution Neural Networks (CNN)

*Convolution Neural Network* (CNN) is one of the most popular deep learning models. This model was the firstly used of digital signal processing. A standard CNN model as shown in Fig. 3, including convolution, pooling, fully connected, and dropout layer [18].

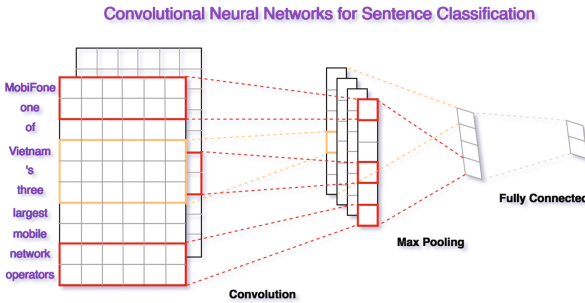


Fig. 3. A standard CNN model [18].

**Convolution Layer:** this layer uses convolution operations to process information by surfing a fixed-size *filter*, or *kernel*, on the input data to obtain more refined data.

**Pooling Layer:** the main purpose of this layer is to perform *pooling* on the vectors outputted from the convolution layer, to allow only the most important vectors to be retained.

**Fully Connected Layer:** in CNN there always has one or several fully connected layer after convolution layer. This is just a typical perceptron-based technique to produce the final output, which is used to re-train again for the whole system in a back propagation manner.

**Dropout:** This is a technique used to prevent overfitting. During the training process, we use a probability  $p$  to randomly prevent some certain weights to be updated.

### 2.2 Word Embedding

**Word Embedding Basic Model:** this model is often used to perform weighting vector. Basically, this is a weight vector, for example, 1-of- $N$  (*one-hot vector*), used to encode

the word in a dictionary of  $M$  words into a vector of length  $M$ . As presented in Fig. 4 is a one-hot vector representing the word Mobifone (supposed that our dictionary has only 5 words: “Viettel”, “Mobifone”, “Vinaphone”, “Sphone”, and “Beeline”).

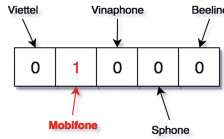


Fig. 4. Representing the “Mobifone” word by one-hot vector.

With such a simple representation, we cannot evaluate the similarity between words since the distance between two vectors always the same (e.g  $\sqrt{2}$  if we use Euclidean distance). Moreover, the dimensionality of the vector space is huge when applied with the real dictionary.

**Word Embedding Model using Word2vec Technique:** *Word2vec* represents the form of a distribution relationship of a word in a dictionary with the rest (known as *Distributed Representation*). As presented in Fig. 5, each word now is represented as a 3-dimensional vector, instead of 5, and each element of the vector is represented by new learning value. In the next section, we will discuss on how to learn those values using neural network on our system.

Viettel	Mobifone	VinaPhone	Sphone	Beeline
0.12	0.41	0.25	0.66	0.34
0.02	0.63	0.33	0.49	0.52
0.01	0.44	0.21	0.23	0.37

Fig. 5. Vectors representing relationships between words.

### 2.3 Long Short-Term Memory (LSTM)

*Long Short-Term Memory* (LSTM) model introduced by Hochreiter and Schmidhuber [21], then improved by Gers et al. [22]. This model is a modification of the model *Recurrent Neural Network* (RNN) [23]. As depicted in Fig. 6, a RNN is arranged in a linear pattern, called *state*, corresponding to a *data entry* of input data. For example, to handle a text document, a state corresponds to a word in the text. Each state received input including the corresponding data entry and the output of the previous state. The output of each state is a weight matrix  $W$ .

In the 1990s, RNNs faced two major challenges of *Vanishing* and *Exploding Gradients*. If the weight  $W$  of each state is too small, the information learned will almost be eliminated when the number of states becomes too large (which is our case of

processing news articles). Conversely, if the weight of  $W$  is large, this will lead to the case known as the *Exploding Gradients* when gradient signal increasingly distracted during training, causing the process not converged.

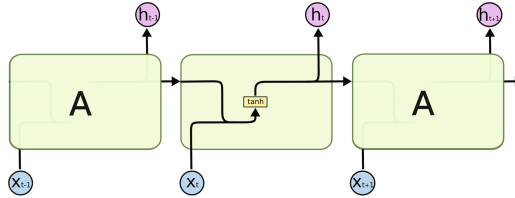


Fig. 6. The standard model with a single roller RNNs layer [23].

LSTM has similar structure to RNN. However, instead of only one layer neural network, a state in a LSTM has 4 layers, as illustrated in Fig. 7. The idea of LSTM is that in each layer there will be a *forget fate*, as illustrated in Fig. 8 to decide whether to allow the previously learned information to be used for the current layer or not. More details on this architecture, interested readers can refer to [22, 24].

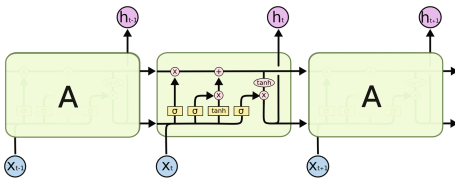


Fig. 7. A LSTM model with 4 layers.

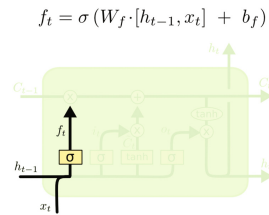


Fig. 8. LSTM forget gate layer.

### 3 The Proposed Deep Architecture

Figure 9 presents an overview of our proposed deep architecture for aspect-level sentiment analysis on news articles. The system includes the following modules.

**Word Embedding Module.** This is a three-layer neural network  $WE$ . The first layer consists of  $M$  input neurons where  $M$  is the number of words in the dictionary. The hidden layer consists of  $K$  nodes, where  $K \ll M$ . The last layer also has  $M$  nodes, and is back propagated to the first layer.

This network will be trained by words in the dictionary. Each word  $w$  will be transmitted to the input layer of  $W$  as an one-hot vector. The network will be trained by a sample of the word  $w'$  similar to  $w$ . Those  $w'$  words can be determined from a predefined domain ontology, or learning from the word co-occurrence from a large corpus of the analyzed domain.

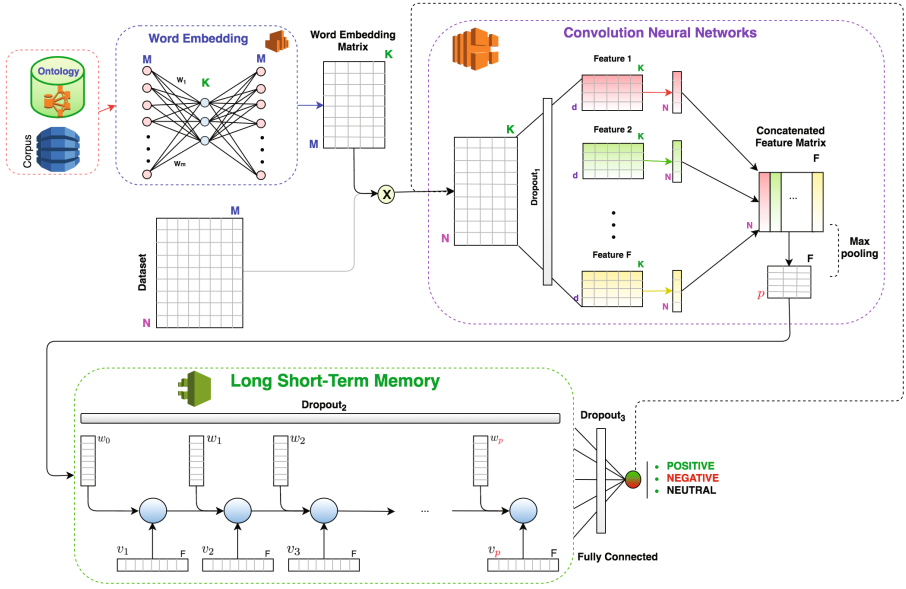


Fig. 9. The overall deep architecture.

After  $WE$  is trained, the weights  $w_{ij}$  on the connections from the  $i^{th}$  node of input layer to  $j^{th}$  node of hidden layer will form a  $W_{M \times K}$  matrix for further usage.

**Training Dataset.** This is a set of collected news articles, each of which was previously labeled as  $\{positive, neutral, negative\}$  over an object  $obj$ . Originally, a document of  $N$  words will be represented as a matrix  $D_{N \times M}$ , in which the  $i^{th}$  row is the one-hot vector of the  $i^{th}$  word of the document. When performing matrix multiplication  $D \times W$ , we get an embedded matrix  $E_{N \times K}$  of the document. Matrix  $E$  will be then used as input for the next *Convolution Neural Network* module.

**Convolution Neural Network.** At this stage, the convolution will be performed between matrix  $E$  with a kernel as a  $F_{d \times K}$  matrix. The meaning of matrix  $F$  is to extract an *abstract feature* based on a hidden analysis of a  $d$ -gram from the original text. There are  $f$  matrices of  $F_{d \times K}$  used to learn  $f$  abstract features. As the convolution of two matrices  $E$  and  $F$  will result in a column matrix of  $N \times 1$ , we will eventually obtain the *convoluted matrix*  $C_{N \times f}$  by combining  $f$  column matrices together.

In the next step, matrix  $C$  will be *pooled* by a pooling window of  $p \times f$ . The significance of this process is to retain the most important  $d$ -gram from  $p$  consecutive  $d$ -gram. Finally, we obtained matrix  $Q_{q \times f}$  where  $q = N/p$ .

At this point, instead of implementing a fully connected layer from the matrix  $Q$  as the CNN-based traditional method, matrix  $Q$  continues to be used as input for the next *LSTM Module*.

**LSTM Module.** This layer consists of  $p$  contiguous states. State  $i$  will receive input as the  $i^{th}$  row on the matrix  $Q$  and *weight vector*  $V_{i-1}$  is the output of the state  $i-1$ .

Thus, each state in the LSTM network is corresponding to a *d-gram*. Due to the characteristics of the LSTM network, if after a number of states, the learning results show that the current data are not very relevant to the learning purpose, the output of the current state will not be passed to the next state. That is, LSTM will start learning again from the next state. As previously discussed, this model is very effective with a kind of article shown in Fig. 2. After handling a number of *d-gram* in paragraphs about the life background of Noo Phuoc Thinh, LSTM network may realize that those contents are not relevant to the previously trained knowledge of the *Mobifone* topic. Thus, the LSTM network may have a chance to restart the learning process with the remaining data.

Vector  $V_q$  will be taken through a final fully connected layer to output the final result of classification (*positive, neutral, negative*). Error from the final result will be propagated back to the beginning layer of the Convolutional Neural Network to continue the training process.

**The Dropout Factors.** As discussed, in order to avoid the overfitting when training data from news articles, we use the dropout factors. As discussed [20], the application of the drop-out can be used for all systems with back propagation learning. As shown in Fig. 9, we used three dropout factors  $p1$ ,  $p2$ ,  $p3$ , respectively for the input layer of the network Convolutional Neural Network, the state transitions on LSTM and the final fully connected layer.

## 4 Experimental Results

### 4.1 Implementation Details

We have implemented a learning system based on the proposed deep architecture and conducted experiments to perform sentiment analysis on the news articles mentioning *Mobifone* subject. In the Word Embedding layer, we have used both *Telco Ontology* and *Telco Corpus* for training. Our *Telco Ontology*, manually constructed, includes the concepts and relationships in the Internet-Telecom domain, as illustrated in Table 1. For the terms that are not captured in our Ontology, we use statistics information from a *Telco Corpus* covering of 12 million articles from electronic news articles, forums, and Facebook social networks. The *Telco Ontology* and *Telco Corpus* are provided by YouNet Media, a company specializing in online data analysis. The original dimension of our one-hot vector is 65000, reduced to 320 after performing Word Embedding.

In Convolution Neural Network layer, we use 64 kernel windows, corresponding to 64 abstract features. The pooling factor is set as 2. All three dropout factors are set at 0.5.

We tested the system with a dataset of 5.000 real articles collected by YouNet Media in the domain of Internet-Telecom. These articles include 401, 743 and 3856 negative, positive and neural articles respectively. The subject to be rated sentiment is *Mobifone*. We applied *k-fold* cross validation strategy with  $k = 5$ .

**Table 1.** Concepts and relations in Internet-Telecom domain.

No	Attribute	Alias keywords	Positive term	Negative term
1	Internet quality	[“chất lượng mạng”/“network quality”, “chất lượng đường truyền”/“line quality”, “nghẽn”/“congestion”, “ngắt kết nối”/“disconnect” ...]	[“nhanh”/“fast”, “mạnh”/“strong”, “khỏe”/“healthy”, “tốt”/“good”, “ổn định”/“stable”...]	[“lạc”/“lag”, “chậm”/“slow”, “kém”/“poor”, “yếu”/“weak”, “đơ”/“pence”, “nhiều”/“noise”, “chập chờn”/“flutter”, “nghẽn”/“congestion” ...]
2	Speed	[“tốc độ”/“speed”, “tải lên”/“upload”, “tải xuống”/“download”, “truyền”/“transmission”, “ngu xuẩn”/“madness”, “như rùa”/“rushing”, “like a turtle” ...]	[“ngon”/“good”, “ầm ầm”/“roaring”, “tốc độ ánh sáng”/“speed of light”, “rầm rầm”/“rumbled”, “ào ào”/“rushing”, “nhanh”/“fast” ...]	[“tệ”/“bad”, “kém”/“poor”, “cùi”/“pulp”, “chậm”/“slow”, “chán”/“boring”, “phàn nàn”/“complaining”, “quay vòng”/“turnaround” ...]
3	Security	[“bảo mật”/“security”, “đánh cắp dữ liệu”/“data theft”, “rò rỉ thông tin”/“leaking information”, “đánh cắp thông tin”/“stolen information” ...]	[“<bảo mật><cao tuyệt đối>”/“<security> <high   absolute>”, “an toàn”/“safe”, “bảo mật hơn”/“more security” ...]	[“bị hack”/“Hacked”, “dở”/“unloading” ...]
	...			

## 4.2 Experimental Results

We applied the deep architecture model to conduct sentiment analysis on the gathered dataset. The original dataset was cloned with variety of variants generation strategies for training. Due to the fact that training data are imbalanced between negative, positive and neutral samples, we use the sampling strategy Smote [25] to re-sample and balance data. Eventually, we obtained two deep learning models, known as *Deep-Unbalanced* and *Deep-Balanced*, corresponding to two cases of applying and not applying the sampling method of Smote.

Results of our models are listed against other methods in Table 2. We divided the data set into the collective experiments *Pos-Dataset* (containing only positive mentions), *Neg-Dataset* (containing only negative mentions), *Neu-Dataset* (containing only neutrals mentions) and *Total-Dataset*. We compare our approach with the traditional SVM classification method using *bag-of-word* approach, along with the two traditional deep learning methods of LSTM and CNN (using Balanced training data).

The results showed that most of the deep learning method have better performance than the traditional SVM method. CNN has achieved also very good performance, but often misclassifies when processing positive articles that the main topic is not



**Table 2.** Experimental results.

	Deep-Unbalanced	Deep-Balanced	SVM	CNN-Balanced	LSTM-Balanced
Pos-Dataset	91.19%	<b>96.17%</b>	76.49%	91.72%	62.37%
Neg-Dataset	78.85%	<b>92.17%</b>	70.49%	90.64%	58.44%
Neu-Dataset	98.82%	<b>99.82%</b>	83.32%	99.06%	66.96%
Total-Model	90.12%	<b>96.52%</b>	80.46	92.12%	64.09%

Mobifone. In particular, the original LSTM suffers from poor performance when applied independently, because news articles often result in very large number of states, heavily affecting the training performance. However, when combining LSTM and CNN as our proposed model, we enjoy very good accuracy, and *Deep-Balanced* model achieved the best accuracy. This demonstrates the advantage of our architecture.

## 5 Conclusion

This paper proposed an architecture to perform deep analysis on news articles sentiment. The neural network models used as sub-modules in this architecture have been previously proposed, including word embedding, CNN and LSTM. Our architecture makes a combination of these models, with careful analysis and explanation of rationales when solving the issues of news articles. We have tested our model with an actual data collection and obtained promising results.

**Acknowledgments.** This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2016-20-36. We are also grateful to YouNet Media for supporting real datasets for our experiment.








## References

1. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003), Sanibel Island, FL, USA, 23–25 October 2003. pp. 70–77 (2003)
2. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, 20–24 May 2003, pp. 519–528 (2003)
3. Padmaja, S., Fatima, S.S.: Opinion mining and sentiment analysis—an assessment of peoples’ belief: a survey. *Int. J. Ad hoc Sensor Ubiquit. Comput.* **4**(1), 21 (2013)
4. Liu, B.: *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael (2012)
5. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. *Comput. Linguist.* **37**(1), 9–27 (2011)

6. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 8–12 May 2007, pp. 171–180 (2007)
7. Li, F., Huang, M., Zhu, X.: Sentiment analysis with global topics and local dependency. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, 11–15 July 2010, Atlanta, Georgia, USA (2010)
8. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, MIT Stata Center, Massachusetts, 9–11 October 2010
9. Sauper, C., Haghighi, A., Barzilay, R.: Content models with attitude. In: Proceedings of the Conference the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 19–24 June 2011, pp. 350–358 (2011)
10. Mukherjee, A., Liu, B.: Aspect extraction through semi-supervised modeling. In: Proceedings of the Conference on the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012, vol. 1, pp. 339–348 (2012). LongPapers
11. Titov, I., McDonald, R.T.: A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008, Columbus, Ohio, USA, 15–20 June 2008, pp. 308–316 (2008)
12. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the Conference on 43rd Annual Meeting of the Association for Computational Linguistics, ACL 2005, University of Michigan, USA, 25–30 June 2005, pp. 115–124 (2005)
13. Le Cun, Y.: Generalization and network design strategies. Technical Report CRG-TR-89-4, University of Toronto Connectionist Research Group, June 1989. A shorter version was published in Pfeifer, Schreter, Fogelman and Steels (eds.) ‘Connectionism in perspective’. Elsevier (1989)
14. Hinton, G.E.: Mapping part-whole hierarchies into connectionist networks. *Artif. Intell.* **46** (1–2), 47–75 (1990)
15. Jenatton, R., Roux, N.L., Bordes, A., Obozinski, G.: A latent factor model for highly multi-relational data. In: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a Meeting, Lake Tahoe, Nevada, United States, 3–6 December 2012, pp. 3176–3184 (2012)
16. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA, 22–27 June 2014, System Demonstrations, pp. 55–60 (2014)
17. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 1631, p. 1642. Citeseer (2013)
18. Kim, Y.: Convolutional neural networks for sentence classification. CoRRabs/1408.5882 (2014)
19. Wang, X., Jiang, W., Luo, Z.: Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In: 26th International Conference on Computational Linguistics, COLING 2016, Proceedings of the Conference: Technical Papers, Osaka, Japan, 11–16 December 2016, pp. 2428–2437 (2016)

20. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
22. Gers, F.: Long short-term memory in recurrent neural networks. Ph.D. thesis, Universitat Hannover (2001)
23. Medsker, L., Jain, L.: *Recurrent Neural Networks: Design and Applications*. CRC Press, Inc., Boca Raton (2001)
24. Understanding LSTM Networks. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Accessed 18 Feb 2017
25. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(5), 855–868 (2009)
26. Batista, G.E.A.P.A., Bazzan, A.L.C., Monard, M.C.: Balancing training data for automated annotation of keywords: a case study. In: *II Brazilian Workshop on Bioinformatics*, Macaé, RJ, Brazil, 3–5 December 2003, pp. 10–18 (2003)

# Sentiment Polarity Detection in Social Networks: An Approach for Asthma Disease Management

Harry Luna-Aveiga<sup>1</sup> , José Medina-Moreira<sup>1</sup> ,  
Katty Lagos-Ortiz<sup>1</sup> , Oscar Apolinario<sup>1</sup> ,  
Mario Andrés Paredes-Valverde<sup>2</sup> ,  
María del Pilar Salas-Zárate<sup>2</sup> , and Rafael Valencia-García<sup>2</sup> 

<sup>1</sup> Universidad de Guayaquil. Cda. Universitaria Salvador Allende,  
Guayaquil, Ecuador

{harry.lunaa, jose.medinamo, katty.lagoso,  
oscar.apolinarioa}@ug.edu.ec

<sup>2</sup> Facultad de Informática, Universidad de Murcia,  
Campus Espinardo, 30100 Murcia, Spain

{marioandres.paredes, mariapilar.salas, valencia}@um.es

**Abstract.** Asthma disease is a serious health problem that affects all age groups. Asthma-related hospitalizations and deaths have declined in some countries. However, the number of patients with symptoms has increased in the last years. Even though asthma patients have contact with health professionals, they must be an active part in treatment team. On the other hand, there has been an exponential growth of information about healthcare and diseases management on social networks such as Twitter. Aiming to benefit from this information, in this work we propose a method for detecting the emotional reaction of patients about asthma domain concepts such as physical activities, drugs, among others. The findings obtained from the analysis of such information can help to other patients to avoid habits that could harm their health. Our proposal was evaluated with a corpus of Twitter messages obtaining a precision of 82.95%, a recall of 82.27%, and F-measure of 82.36% in sentiment polarity identification.

**Keywords:** Ontology · Asthma · Twitter · Sentiment polarity

## 1 Introduction

Asthma disease is a global health problem that affects all age groups, ranging from 1% to 21% in adults [1] and with up to 20% of children [2]. This disease impacts not only patients, but also their families, as well as healthcare systems and society [3]. Although asthma-related hospitalizations and deaths have declined in some countries [4], the number of patients with day-to-day symptoms has increased by almost 30% in the past 20 years [5]. Despite the fact that asthma patients have periodic contact with health professionals, in asthma disease management the patients should not be seen as objects for treatment but rather as active participants in a treatment team [6].

There has been an exponential growth of information about healthcare and diseases management on social media such as forums, blogs, microblogs, and social networks. One of the most used social networks is Twitter, which has become a powerful tool for disseminating experiences and fostering diseases self-management conversations. Aiming to benefit from this information, several organizations apply sentiment analysis techniques to determine the attitude or emotional reaction of patients to topics such as drugs, guidelines, healthcare services, among others. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language [7].

In this work, we propose a method for detecting the emotional reaction of asthma disease patients about risk factors, physical activities, among other concepts. In this way, the findings obtained can help to other patients to avoid habits that could harm their health. Furthermore, our approach can help to raise awareness about asthma, thus motivating additional help-seeking behavior. The approach here presented relies in two main technologies. On the one hand, it takes advantage of Semantic Web technologies, more specifically from ontologies for representing the asthma's domain. An ontology is a formal and explicit specification of a shared conceptualization [8]. This technology has been applied in different domain such as natural language interfaces [9], cloud computing [10], recommender systems [11], and finances [12], among others. On the other hand, it uses SentiWordNet [13] to determine the polarity of asthma domain concepts contained within Twitter messages based on the words close to them.

The remainder of this paper is structured as follows. Next section describes the most relevant works about opinion mining in medical domain. Section 3 describes the way our approach is decomposed into four *constituent* modules and the ways these elements interact with each other. Section 4 presents a discussion about the evaluation results obtained by our approach. Finally, our conclusions and future work are summarized.

## 2 State of the Art

Sentiment polarity identification in social media have attracted increasing attention in medical and healthcare domains. For instance, in [14] the authors present their effort to examine Twitter conversations to know how breast cancer screening guidelines changes are accepted and implemented into practice by patients and providers. The authors employed standard descriptive statistics to summarize the results. The findings obtained suggest that it is necessary to disseminate accurate information regarding breast cancer prevention modalities. In [15] the authors present a method for polarity classification of patient's experiences about drugs. The method proposed consists of two steps. Firstly, a knowledge base of polar facts is generated from Linked Data sources. Secondly, the method exploits the knowledge base for polarity classification of a dataset collected from websites for reviewing drugs. In [16] the authors presented an approach to identify influential online health community (OHC) users that through their online activities are able to affect the emotion of other community members. Furthermore, the authors proposed a novel metric that directly measures a user's ability to affect the sentiment of others. For this purpose, text mining and sentiment analysis

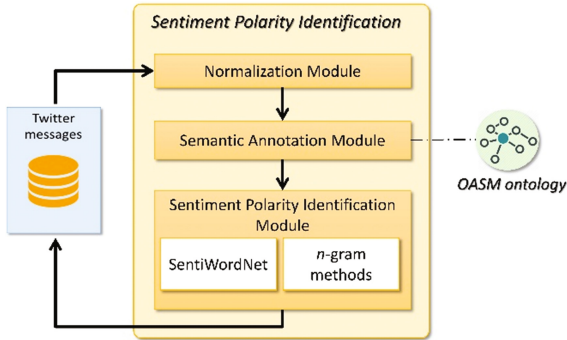
technologies were used. Finally, the authors collected a corpus of 468000 posts, which were manually labeled to carry out their validation experiments. On the other hand, Segura et al. [17] developed a Spanish corpus of users comments about drugs. Also, they proposed a system for detecting drug adversities from Spanish health social media. Subsequently, this system was validated by using the corpus collected obtaining good results. In [18], the authors proposed a sentiment analysis approach to determine the positive or negative aspects of health care domain. For instance, to determine whether the hospital facilities were clean. This approach consists of two main components: a pre-processing module, and a sentiment classification module. The first module splits patients' comments into manageable units aiming to build a formal representation of the data. The second module performs the training of different machine learning algorithm namely, Naive Bayes, Decision trees, and Support Vector Machine, in order to allow determining the sentiment polarity of new comments. Finally, Arco et al. [19] presented a data resource for opinion mining known as COPOS (Corpus Of Patient Opinions in Spanish). To validate the corpus, the authors performed some experiments based on the semantic orientation and machine learning approaches. More specifically, the SVM algorithm and the TF-IDF feature extraction methods were used for supervised machine learning. Meanwhile, the iSOL lexicon was used for the semantic orientation approach.

Based on the study presented above, our approach has some relevant and unique aspects: (1) it is focused on the asthma disease domain, which has not yet been explored, (2) one of its main goals is to facilitate the self-management of health through the analysis of opinions in social networks, and (3) it performs the semantic annotation of Twitter messages by using an ontology for asthma disease management, which has been also proposed in this work.

### 3 Sentiment Polarity Identification

Figure 1 provides a graphical representation of our approach. As can be seen, it is decomposed into four constituent elements: (1) normalization module, (2) semantic annotation module, (3) the ontology for the asthma self-management, and (4) the sentiment polarity identification module.

In a nutshell, our approach works as follows. The first module normalizes the tweets collected, i.e., it removes special characters, expands abbreviations and correct spelling errors contained in them. The second module processes the tweets to obtain all occurrences of concepts that are of special interest in asthma disease self-management. To this end, this second module uses the ontology for the asthma self-management called OASM (Ontology for Asthma Self-Management), which describes concepts such symptoms, risk factors, drugs and physical activities, among others. Finally, the fourth module determines the sentiment polarity of each concept by summing the negative, positive and neutral polarity of each word close to the concept. This polarity is given by SentiWordNet [13] a lexical resource for opinion mining. Next sections provide a wider description of each component mentioned above.



**Fig. 1.** Architecture

### 3.1 Normalization Module

Twitter messages vary in content and composition, often containing non-standard words, ungrammatical sentence structures and domain-specific terms, also known as jargon. This phenomenon reduces the accuracy in many natural language processing tasks [20]. Aiming to address this issue, the present module focuses on normalize Twitter messages, i.e., normalize non-standard words to their canonical form. Furthermore, it removes characters that do not contribute to the meaning of the tweet. More specifically, this module performs next tasks:

1. Removing special characters. Most tweets contain special characters that do not contribute to the goal of this work. Hence, this module removes all @ that are part of mentions and replies to other tweets, # corresponding to hashtags, and URLs.
2. Hashtag normalization. Often, hashtags are written in CamelCase, i.e., they are composed of words of phrases such that each word in the middle of the hashtag begin with a capital letter, with no intervening punctuation or spaces.
3. Expand abbreviations. Because of Twitter limits Tweet length to 140 characters, users must use informal abbreviations (e.g., *bc* for because and *tmrw* for tomorrow) as well as phonetic substitutions (e.g., *b4* for before and *4eva* for forever). The use of this kind of terms makes difficult the understanding of the Tweets by means of the natural language processing techniques implemented in this work. To deal with this problem, this module employs an SMS (Short Message Service) dictionary to expand the abbreviations frequently used by social networks' users.
4. Correction of spelling errors. This task aims to correct orthographical errors. To this end, this module uses Hunspell [21] a spell checker and morphological analyzer.

### 3.2 Semantic Annotation Module

Our approach needs to know all those concepts that are of special interest in asthma's self-management. Hence, in this work we design an ontology called OASM that describes concepts such as risk factors, drugs, symptoms, and other concepts concerning asthma care. This ontology is based on already available ontologies for disease

management. More specifically, it includes concepts from the Asthma Ontology (AO), a biomedical ontology used for modelling asthma related data in routine clinical databases. AO was developed as part of the MOCHA (Model for Child Health Appraised) project [22]. The AO ontology describes concepts related to diagnosis, symptoms, therapy and process of care. The Fig. 2 shows an extract of the ontology.

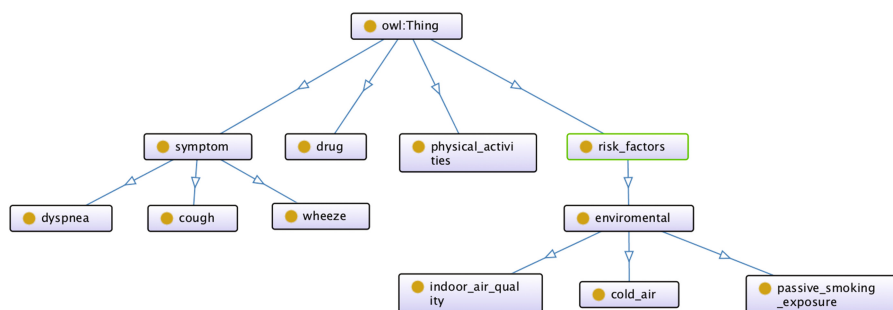


Fig. 2. Excerpt from OASM ontology

The classes contained in the basic hierarchy of OASM are:

- **Symptom.** This class contains a taxonomy of asthma's symptoms. A symptom is a phenome accompanying something, in this case asthma disease, and is regarded as evidence of its existence [23]. Among the symptoms described by this OASM are chest tightness, airway inflammation, shortness of breath, cough, wheeze, dyspnea, among others.
- **Risk factors.** According to the WHO (World Health Organization) a risk factor is any attribute, characteristic or exposure of an individual that increases the likelihood of developing a disease or injury [24]. Some of the risk factors described by this ontology are allergens such as dust mite, mold, house dust; environmental conditions such as cold air, passive smoking exposure, indoor air quality; to mention but a few.
- **Drug.** This class represents asthma medication including those drugs that prevent asthma attacks as well as quick-relief medications, also known as rescue medications. Furthermore, it represents all drugs to which many people with asthma have sensitivities. For instance, some common medications known to trigger symptoms of asthma are aspirins and NSAIDS (non-steroidal anti-inflammatory drugs) such as ibuprofen and naproxen.
- **Physical activities.** This class describes exercises and sports that can have a positive effect on asthma symptoms. Also, it describes sports that can affect the asthma symptoms because they require continuous exertion or because the cold, dry air present in the places where they are practiced. Examples of this kind of sports are running, soccer, basketball, among others.



Based on the ontology described above, this module process all tweets to recognized all domain's concepts. This process is carried out using the GATE framework [25], a natural language processing tool for developing software components that process human language.

### 3.3 Sentiment Polarity Identification Module

This module identifies the sentiment polarity of the concepts recognized by the previous module considering the words close to them. To this end, this module implements the  $n$ -gram method, which has been applied in research works such as [12, 26, 27]. In the computational linguistic domain, an  $n$ -gram is defined as a contiguous sequence of  $n$  items from a given sequence of text, where items can be phoneme, words, or other elements according to the application. Specifically, our approach considers the words (from 2 to 6) after, before, and around the concept using the  $n$ -gram after,  $n$ -gram before, and  $n$ -gram around methods, respectively.

The sentiment polarity of a concept is determined by the sum of the polarity of all words close to it, which were obtained by the  $n$ -gram method. The polarity of each word is given by SentiWordNet (hereafter referred as SWN), a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. SWN is the result of automatically annotating all WordNet [28] *synsets* according to their degrees of positivity, negativity, and neutrality. A WordNet *synset* is a set of cognitive synonyms consisting of nouns, verbs, adjectives and adverbs, each expressing a distinct concept. The positive, negative and neutral polarity degree of each *synset* corresponds to a numeric value between 0 and 1, being the sum of all them equal to 1. In this way, the positive, neutral and negative polarity of each concept is calculated through next equations:

$$PosPol(a_i) = \sum_{w \in wa_i} PosPolSWN_w \quad (1)$$

$$NegPol(a_i) = \sum_{w \in wa_i} NegPolSWN_w \quad (2)$$

$$NeuPol(a_i) = \sum_{w \in wa_i} NeuPolSWN_w \quad (3)$$

Where  $a$  represents the concept identified;  $wa_i$  is the set of words close to the concept, which were obtained by the  $n$ -gram method; and  $PosPolSWN$ ,  $NegPolSWN$  and  $NeuPolSWN$  represent the polarity (positive, negative and neutral, respectively) given by SWN to each word ( $w$ ). Once these polarities are computed, the polarity with the highest score determines the attitude of asthma disease patients with respect to the specific concept or the contextual polarity or emotional reaction to the concept.

As was previously mentioned, each WordNet *synset* can consist of nouns, verbs, adjectives or adverbs. In other words, a *synset* contains one or more words, which means that multiple words can represent the same sense, or meaning. Furthermore, a word can appear in multiple *synsets*. For instance, it can belong to 8 *synsets*, 3 noun senses, 3 verb senses, and 2 adjective senses. Based on this understanding, it is

necessary to disambiguate all words selected by the n-gram method, i.e., it is necessary to interpret the author's intended use of a word within the Twitter message. With this purpose, the present module implements Babelify [29], a unified, multilingual, graph-based approach to Word Sense Disambiguation.

## 4 Evaluation

### 4.1 Procedure

The main objective of this evaluation was to know how successful our approach was to identify the sentiment polarity of an asthma self-management concept contained within Twitter messages. For this purpose, we performed a prospective analysis of 600 tweets containing one or more of the concepts described in the OASM ontology. The tweets were obtained using the Twitter API [30]. To ensure the quality of the corpus, a group of healthcare experts analyzed and tagged all tweets collected. More specifically, the experts detected the asthma self-management concepts contained in the tweet and classified each one as positive, negative, or neutral. Since our method is based on semantic orientation approach, the main objective of tagging these concepts is to obtain the baseline results to evaluate our proposed method.

Once corpus was collected and tagged, we evaluated the effectiveness of our approach by using three metrics namely, precision, recall [31], and F-measure [32]. These metrics are defined by the following equations:

$$Precision = \frac{True\ positives}{True\ positives + False\ positives} \quad (4)$$

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives} \quad (5)$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

In this evaluation, we used three classes corresponding the sentiment polarity of concepts, i.e., the positive, negative and neutral classes. In this way, the precision, recall and F-measure scores of the positive class are computed as follows. *True positives* are the concepts correctly classified as positive, *False negatives* are the positive concepts that were classified as negative or neutral, and *False positives* are the negative or neutral concepts that were classified as positives.

In a multiclass classification, such as the performed in this work, the precision, recall, and F-measure scores are computed for each class. Hence, to obtain a complete evaluation of the system, the evaluation results of each class must be combined. For this purpose, the macro-average measure is used [33]. This metric is the arithmetic mean of precision, recall and F-measure, where the quotient is the number of classes (C) with which the problem counts. The Macro-Precision and Macro-Recall equations are presented below:

$$\text{Macro - Precision} = \frac{\sum_{i=1}^{|C|} \text{Precision}}{|C|} \quad (7)$$

$$\text{Macro - Recall} = \frac{\sum_{i=1}^{|C|} \text{Recall}}{|C|} \quad (8)$$

Regarding the macro of the F-measure, it is the harmonic mean of macro-Precision and macro-Recall.

## 4.2 Results

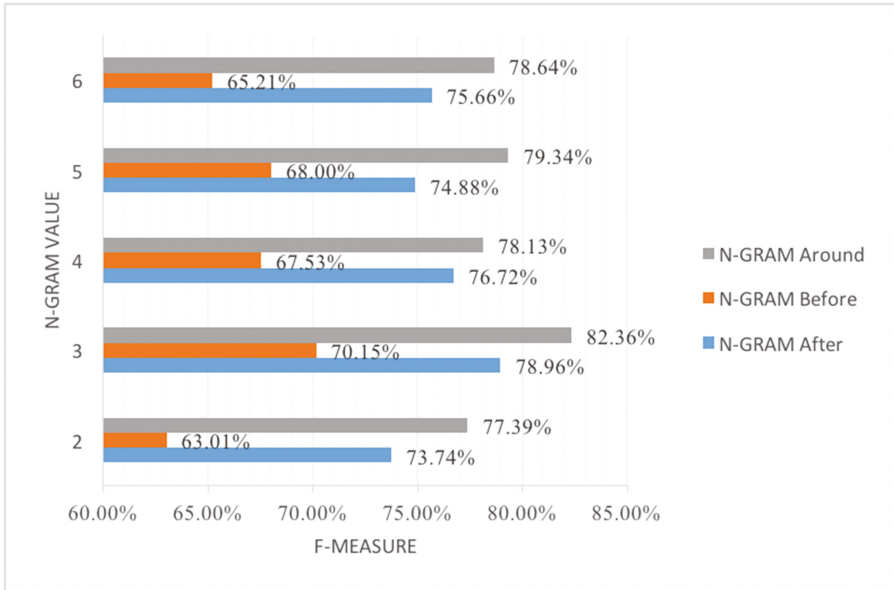
Table 1 presents the macro-average results (Precision (P), Recall (R), and F-measure (F1)) obtained for the different  $n$ -gram methods. As can be seen, the first column represents the number of words considered by the  $n$ -gram method, in this case 2 to 6. The other columns present the P, R, and F1 values for each  $n$ -gram method.

Furthermore, Table 1 shows that the three  $n$ -gram methods obtained the best results when three words close to the concept are considered.  $N$ -gram after obtained a precision of 79.23%, a recall of 78.87%, and an F-measure of 78.96%.  $Then$ -gram before method obtained a precision of 70.47%, a recall of 70.07%, and an F-measure of 70.15%. Finally,  $then$ -gram around method obtained a precision of 82.95%, a recall of 82.27% an F-measure of 82.36%. Meanwhile, the worst results for the three methods were obtained when only two words close to the concept were considered. The  $n$ -gram after method obtained a precision of 74.19%, a recall of 73.60%, and an F-measure of 73.74%.  $Then$ -gram before method obtained a precision of 64.29%, a recall of 62.80%, and an F-measure of 63.01%. The  $n$ -gram around method obtained a precision of 78.09%, a recall of 77.27%, and an F-measure of 77.39%.

**Table 1.** Evaluation results

$n$ -gram	$n$ -gram before			$n$ -gram after			$n$ -gram around		
	P	R	F1	P	R	F1	P	R	F1
2	74.19	73.60	73.74	64.29	62.80	63.01	78.09	77.27	77.39
3	79.23	78.87	78.96	70.47	70.07	70.15	82.95	82.27	82.36
4	77.12	76.60	76.72	68.11	67.47	67.53	78.99	78.00	78.13
5	75.34	74.73	74.88	68.62	67.93	68.00	80.01	79.21	79.34
6	76.06	75.53	75.66	66.06	65.13	65.21	79.00	78.60	78.64

Figure 1 shows a graphical representation of evaluation results. In such figure, we can perceive that the  $n$ -gram around method obtained better results than the  $n$ -gram after and  $n$ -gram before methods. Specifically, the best result (an F-measure of 82.36%) was obtained with an  $n$ -gram value of 3. Based on these results, we can conclude that the three words preceding and following the concept allow identifying better its sentiment polarity (Fig. 3).



**Fig. 3.** Sentiment polarity identification by means of  $n$ -gram methods.

General results are promising and confirm the effectiveness of our method for identifying the sentiment polarity of asthma self-management concepts contained within Twitter messages. However, we considered that our approach can be improved in two main ways. First, our approach depends on a sentiment lexicon, in this case the SentiWordNet lexicon. Although this lexicon has been successfully used in sentiment classification tasks, we are considering the use of a health domain lexicon that allow us to determine in a better way the polarity of words occurring in the context considered in this paper. Second, the normalization process of Twitter messages represents a difficult task due to the use of ungrammatical sentences structures as well as the use of non-standard words. Therefore, we would like to explore more methods that help to normalize the Twitter messages prior to the application of the  $n$ -gram methods.

## 5 Conclusions and Future Work

Twitter is an essential part of the dissemination of diseases self-management, but little work has been done examining this topic in the asthma disease management. Consequently, we propose an approach based on sentiment analysis for determining the attitude of asthma disease patients with respect to several domain concepts. Furthermore, this approach takes advantages of Semantic Web technologies, more specifically ontologies to model the domain under consideration. Our approach obtained encouraging results with a precision of 82.95%, a recall of 82.27%, and F-measure of 82.36%. Furthermore, the evaluation results show the effectiveness of the  $n$ -gramaround method for sentiment polarity identification on the health domain.

We think that our approach could be coupled with methods for detecting the figurative language such as sarcasm and irony. Sarcastic and ironic comments are commonly written on the social networks. This phenomenon requires being considered to determine the correct polarity of the domain concepts due to it can play the role of polarity reverse. Also, we think that our approach can be adapted to other languages. However, this adaptation process requires resources and NLP tools that are difficult to find, or that are not yet mature. Therefore, we will attempt to apply our approach to languages such as Spanish, French, Arabic, to verify the effectiveness of our method. On the other hand, we plan to use the corpus collected in this work to develop a supervised machine learning system for the automatic detection of emotional reaction of patients about asthma domain, and then comparing the results with those obtained in this work. Finally, despite our approach deals with textual error phenomena by means of a normalization process based on techniques such as hashtag normalization, expanding abbreviations and correction of spelling errors, we plan to extend this method to deal with problems such as foreign terms, free inflections and character repetition.

**Acknowledgements.** This work has been funded by the Universidad de Guayaquil (Ecuador) through the project entitled “Tecnologías inteligentes para la autogestión de la salud”. María del Pilar Salas-Zárate and Mario Andrés Paredes-Valverde are supported by the National Council of Science and Technology (CONACYT), the Secretariat of Public Education (SEP) and the Mexican government. Finally, this work has been also supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through project KBS4FIA (TIN2016-76323-R).

## References

1. To, T., et al.: Global asthma prevalence in adults: findings from the cross-sectional world health survey. *BMC Pub. Health* **12**(1), 204 (2012)
2. Lai, C., Beasley, R., Crane, J., Foliaki, S., Shah, J., Weiland, S.: Global variation in the prevalence and severity of asthma symptoms: phase three of the International Study of Asthma and Allergies in Childhood (ISAAC). *Thorax* **64**(6), 476–83 (2009)
3. Reddel, H.K., et al.: A summary of the new GINA strategy: a roadmap to asthma control. *Eur. Respir. J.* **46**(3), 622–639 (2015)
4. Lozano, R., et al.: Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *Lancet* **380**(9859), 2095–2128 (2013)
5. Vos, T., et al.: Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *Lancet* **380**(9859), 2163–2196 (2012)
6. Lahdensuo, A.: Guided self management of asthma—how to do it. *BMJ* **319**(7212), 759–760 (1999)
7. Liu, B.: Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **5**(1), 1–167 (2012)
8. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. *Data Knowl. Eng.* **25**(1), 161–197 (1998)

9. Paredes-Valverde, M.A., Valencia-García, R., Rodríguez-García, M.Á., Colomo-Palacios, R., Alor-Hernández, G.: A semantic-based approach for querying linked data using natural language. *J. Inf. Sci.* (2015). p. 0165551515616311
10. Rodríguez-García, M.Á., Valencia-García, R., García-Sánchez, F., Samper-Zapater, J.J.: Creating a semantically-enhanced cloud services environment through ontology evolution. *Future Gener. Comput. Syst.* **32**, 295–306 (2014)
11. Carrer-Neto, W., Hernández-Alcaraz, M.L., Valencia-García, R., García-Sánchez, F.: Social knowledge-based recommender system. Application to the movies domain. *Expert Syst. Appl.* **39**(12), 10990–11000 (2012)
12. del Pilar Salas-Zárate, M., Valencia-García, R., Ruiz-Martínez, A., Colomo-Palacios, R.: Feature-based opinion mining in financial news: an ontology-driven approach. *J. Inf. Sci.* (2016). p. 0165551516645528
13. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *LREC* **10**, 2200–2204 (2010)
14. Nastasi, A., Bryant, T., Canner, J.K., Dredze, M., Camp, M.S., Nagarajan, N.: Breast cancer screening and social media: a content analysis of evidence use and guideline opinions on twitter. *J. Cancer Educ.* **59**, 1–8 (2017)
15. Noferesti, S., Shamsfard, M.: Using linked data for polarity classification of patients' experiences. *J. Biomed. Inform.* **57**, 6–19 (2015)
16. Zhao, K., Yen, J., Greer, G., Qiu, B., Mitra, P., Portier, K.: Finding influential users of online health communities: a new metric based on sentiment influence. *J. Am. Med. Inform. Assoc.* **21**(e2), e212–e218 (2014)
17. Segura-Bedmar, I., Revert, R., Martínez, P.: Detecting drugs and adverse events from Spanish health social media streams. In: *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL*, pp. 106–115 (2014)
18. Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L.: Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J. Med. Internet Res.* **15**(11), e239 (2013)
19. del Arco, F.M.P., Valdivia, M.T.M., Zafra, S.M.J., González, M.D.M., Cámara, E.M.: COPOS: corpus of patient opinions in spanish. application of sentiment analysis techniques. *Procesamiento del Lenguaje Natural* **57**, 83–90 (2016)
20. Liu, F., Weng, F., Wang, B., Liu, Y.: Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers*, vol. 2, pp. 71–76 (2011)
21. Hunspell: About. <http://hunspell.github.io/>. Accessed 25 Feb 2017
22. Blair, M., Blair, M.: 3. N. Regular workshop: Finding and Implementing 'Best' Child Primary Health Care-Key Themes in the new MOCHA Project
23. Patil, J.K., Kumar, R.: Advances in image processing for detection of plant diseases. *J. Adv. Bioinform. Appl. Res.* **2**(2), 135–141 (2011)
24. WHO|Risk factors: WHO. [http://www.who.int/topics/risk\\_factors/en/](http://www.who.int/topics/risk_factors/en/). Accessed 26 Feb 2017
25. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.: Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput. Biol.* **9**(2), e1002854 (2013)
26. Penalver-Martinez, I., et al.: Feature-based opinion mining through ontologies. *Expert Syst. Appl.* **41**(13), 5995–6008 (2014)
27. del Pilar Salas-Zárate, M., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M.Á., Valencia-García, R.: Sentiment analysis on tweets about diabetes: an aspect-level approach. *Comput. Math. Meth. Med.* **2017** (2017). Article ID 5140631

28. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
29. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. *Trans. Assoc. Comput. Linguist. TACL* **2**, 231–244 (2014)
30. Makice, K.: *Twitter API: Up and Running: Learn How To Build Applications with the Twitter API*. O'Reilly Media, Inc., Sebastopol (2009)
31. Clarke, S.J., Willett, P.: Estimating the recall performance of Web search engines. In: *Aslib Proceedings*, vol. 49, pp. 184–189 (1997)
32. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42–49 (1999)
33. Lewis, D.D.: *Representation and Learning in Information Retrieval*. University of Massachusetts, Amherst (1992)

# An Overview of Information Discovery Using Latent Semantic Indexing

Roger Bradford<sup>(✉)</sup> 

Maxim Analytics, Reston, VA 20190, USA  
rogerbbradford@gmail.com

**Abstract.** In recent years there has been a dramatic increase in the size of information collections of importance. At the same time, there has been a growing interest in extracting as much useful information as possible from such collections. These trends place significant demands on modern information retrieval systems. In particular there is a great need for tools that can support discovery of new and useful information. The technique of latent semantic indexing (LSI) has a number of attributes that make it particularly well-adapted to information discovery applications. This paper provides an overview of LSI-based techniques that have been successfully employed in facilitating discovery in practical applications. The techniques range from user aids to state-of-the-art discovery methods.

**Keywords:** Information discovery · Latent semantic indexing · LSI · Latent semantic analysis · LSA · Query reformulation · Novelty detection

## 1 Introduction

One of the most significant trends in commercial and government information systems in recent years has been the desire to extract actionable information from increasingly large data collections.

As collections become larger, however, the problem of efficiently extracting useful information from them becomes more complex. Some of the more vexing issues that arise are:

- In many collections, the proportion of information that is relevant to a given problem or to a given user's interests may be very low.
- For text, as the size of collections grows, the rapid increase in the number of terms makes mismatch between user terminology in queries and author terminology in documents increasingly likely.
- The larger the collection, the more variants of important object descriptors, such as people's names, are likely to occur.
- In many large collections there are great amounts of redundant data, which hinders finding new information of interest.
- In large collections it is difficult to decide which of the enormous number of relationships among items are worthy of investigation.



In recent years there has been a growing awareness that, in dealing with big data issues, information *retrieval* tools alone are inadequate [1]. Both analytic software and human users must be provided with tools that aid in the *discovery* of key information, including relevant entities, new relationships, and unexpected contexts. Fortunately, there are emerging tools that can significantly improve the efficiency of carrying out such discovery activities. Some of the greatest successes in this area have come from exploiting the technique of latent semantic indexing (LSI).

This paper presents examples of how LSI can be used to provide discovery functionality in modern information systems. In order to bound the discussion, this paper deals with applications involving human users and textual data. However, the techniques described are equally applicable for situations involving automated agents and for other types of data [2].

## 2 Latent Semantic Indexing

LSI is an information organization and analysis tool that has wide applicability. LSI (as applied to text) accepts as input a collection of documents and produces as output a high-dimensional vector space.

As applied to a collection of documents, the algorithm consists of the following primary steps [3]:

1. A matrix  $A$  is formed, wherein each row corresponds to a term that appears in the documents, and each column corresponds to a document. Each element  $a_{m,n}$  in the matrix corresponds to the number of times that the term  $m$  occurs in document  $n$ .
2. Local and global term weighting is applied to the entries in the term-document matrix. This weighting improves the ability to distinguish among items in the space produced. Some very common words such as *and*, *the*, etc. typically are deleted entirely (i.e., treated as stop words).
3. Singular value decomposition (SVD) is used to reduce this matrix to a product of three matrices:

$$A = U \Sigma V^T$$

$\Sigma$  is a diagonal matrix whose elements are the singular values of  $A$  (the non-negative square roots of the eigenvalues of  $AA^T$ ).

4. Dimensionality is reduced by deleting all but the  $k$  largest values of  $\Sigma$ , together with the corresponding columns in  $U$  and  $V$ , yielding an approximation of  $A$ :

$$A_k = U_k \Sigma_k V_k^T$$

which is the best rank- $k$  approximation to  $A$  in a least-squares sense.

5. This truncation process provides the basis for generating a  $k$ -dimensional vector space. Both terms and documents are represented by  $k$ -dimensional vectors in this vector space.
6. New documents (e.g., queries) and new terms can be represented in the space by a process known as folding-in, which extrapolates from known vectors [4].

7. The similarity of any two objects represented in the space is reflected by the proximity of their representation vectors, generally using a cosine measure.

Extensive experimentation has shown that proximity of objects in such a space is an effective surrogate for conceptual similarity in many applications [5].

### 3 Relevant Work

Legal discovery is the largest commercial application of LSI [6]. In the U.S., the term *discovery* has a specific meaning in the legal domain: the pre-trial phase in a lawsuit in which each party can obtain evidence from the opposing party. Historically, identification of relevant documents during legal discovery was an intensely manual process based on Boolean retrieval. In recent years, the rapid growth in the size of litigation-related document collections has led to application of more capable techniques [7]. LSI has become by far the dominant technical approach used in this area.

LSI also has been used to:

- Discover implicit relationships among individuals in e-mail collections [8].
- Detect anomalies in graphs of relationships [9].
- Automatically populate ontologies used in discovery efforts [10].
- Extract implicit geographic information from textual data [11].
- Automatically discover semantic links among geospatial data resources [12].
- Discover structural relationships among patents for cross-domain design [13].
- Leverage user discovery experiences in geographic information portals [14].
- Discover architectural information in software product audits [15].
- Identify relationships across media types [16].
- Automatically identify regions of interest in images [17].
- Uncover computer network attacks [18].
- Discover insider threats [19].
- Uncover scenarios in large collections of unstructured data [20].
- Discover similarities over large scale heterogeneous data [21].
- Discover web services of interest [22].
- Uncover software vulnerabilities [23].
- Discover relationships between databases [24].
- Detect implicit information in social media [25].
- Support discovery processes incorporating common sense information [26].
- Facilitate discovery in adversarial situations (e.g., where aliases are used) [27].

The most rapidly growing application area for LSI is in the biomedical field.<sup>1</sup> In that discipline, LSI has been applied to discovery of: relationships between gene

---

<sup>1</sup> A review of major online sources (Science Direct, Springer Link, Google Scholar, and the digital libraries of the IEEE and ACM) indicates that publications regarding biomedical applications of LSI were negligible prior to 2000, grew linearly between then and 2007, surged in 2008, and have been growing at a rate of 10–15% per year since then.

sequences and amino acids [28], functional relationships between genes [29], relations between genes and diseases [30], gene-gene interactions [31], gene co-expression linkages [32], linkages between medical publications and annotations in the Gene Ontology data collection [33], transcription factors for genes [34], functional cohesion of gene sets [35], regulatory factors [36], protein-protein interactions [37], and potential drug targets [38], as well as other items.

## 4 Example Applications

The complete range of techniques through which LSI is applied in information discovery is much too large to cover in this paper. The intent here is to provide an overview of key relevant aspects of information discovery tradecraft. Several examples are described in order to provide insight into methods and best practices. As a unifying measure, the discussion focuses on a single application area - analysis by users of unstructured data. The defining element of each example is the identification of new, relevant information in either of the following cases:

- When the user does not have sufficient information to formulate an effective query.
- When identification of the information of interest using conventional tools (e.g., Boolean retrieval systems) would be extremely laborious.

The discussion generally moves in a direction of increasing complexity, from simple user aids through examples of the current state of the art in LSI-based discovery applications.

### 4.1 Query Expansion

One problem immediately encountered when attempting to extract information from large text collections is the huge number of terms encountered. Figure 1 shows the number of unique terms indexed in 64 LSI spaces used in two dozen typical modern English-language LSI applications.<sup>2</sup> The numbers are much larger than the number of unique English words present in the collections because of the occurrence of technical terms, acronyms, proper names, equipment designators, etc.

With that broad a range of terminology, requiring a user to think of all the ways in which a concept of interest might have been expressed is unreasonable. Tools must be provided to aid users in formulating queries. This is an application where the conceptual mapping capability of LSI (described in Sect. 2) has been applied with great success. Two methods of expanding queries using LSI are described below as examples.

---

<sup>2</sup> These figures are from commercial and government applications worked on by the author between 2005 and 2013. These applications primarily involved conceptual retrieval, text clustering, and/or document categorization tasks. Most had at least some focus on new information discovery.

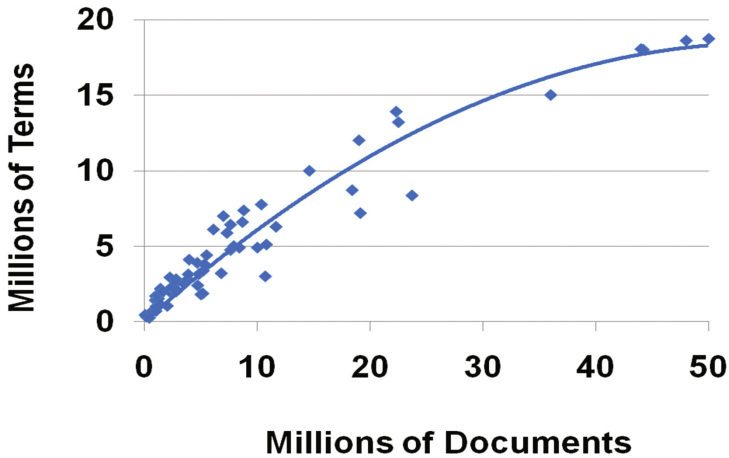


Fig. 1. Terms vs. documents in representative LSI applications

**Discovering Relevant Terms and Documents for Query Expansion.** A standard operation in an LSI space is to list terms that are semantically most closely associated with a query, in order of the strength of the association. This capability can be of great use in expanding queries. For example, in one LSI index of nine million documents, the terms most closely associated with a query on social unrest were:

- discontent
- protest
- demonstrated
- riot
- barricades
- stand-off.

A user might readily have thought of the first four of these as query terms but might well not have thought immediately of the last two. Identifying useful additional query terms in this manner is a feature of many modern LSI applications.

Using LSI, this type of query expansion also can be carried out at the document level, iteratively concatenating retrieved documents with a user's original query. In large collections, such query concatenation can more than double recall while maintaining good precision [39].

**Discovering Name Variants for Query Expansion.** In many applications, names of persons are particularly important query terms. In large collections, there may be from tens to over 100 variants of a given name due to contractions, nicknames, varying completeness, name order, misspellings, typographical errors, phonetic renderings, transliteration differences and varying associated titles. Within an LSI space, variants of names are automatically closely associated, due to their similar contexts. This fact has been applied with great success in identifying name variants in large document collections [40]. The approach is to generate large numbers of candidate name variants

based on standard techniques (Levenshtein Metric, Dice Coefficient, etc.) and then select the most likely correct ones based on similarity of their LSI contexts. For example, in a collection of less than one million news articles, this approach identified 19 variants of the name and associated titles of former Columbian president Alvaro Uribe Velez, as shown in Table 1.

**Table 1.** Variants of *Alvaro Uribe Velez* in less than one million documents

President Alvaro Uribe Velez	Leader Alvaro Uribe Velez	Doctor Alvaro Uribe
President Elect Alvaro Uribe Velez	President Elect Alvaro Uribe	Presidents Alvaro Uribe
Alvaro Uribe Velez	President Alvaro Uribe	Uribe Velez
Doctor Alvaro Uribe Velez	President Uribe Velez	Velez Uribe
Governor Alvaro Uribe Velez	Governor Alvaro Uribe	Mr. Uribe Velez
Mr. Alvaro Uribe Velez	Alvaro Uribe	President Uribe
Senator Alvaro Uribe Velez		

This is a good example of the second type of data exploration described above – the user would expect that variants of the name exist; however, it would require considerable thought and significant trial and error to find those variants using conventional tools, such as Boolean queries. This type of query expansion also is applied for other types of entities, such as organizations.

### 4.2 Topic Discovery via Clustering

Document clustering is a powerful approach for providing a user with a quick overview of the contents of a collection of documents. For example, Fig. 2 shows the results of automated clustering and labeling of 5,000 survey forms collected by a large hotel chain.

Similarity comparisons in an LSI space can be employed to produce topic clusters such as these, using a wide variety of clustering techniques. Retrieving the closest



**Fig. 2.** Example of LSI cluster generation and labeling

terms to the cluster centroids provides meaningful labels for the clusters, as shown in the figure. This technique primarily is used to provide a quick overview of data. However, there is opportunity for discovery when an unexpected cluster occurs.

### 4.3 Novelty Detection

One difficulty in working with unstructured information is that in many cases there is a great deal of redundancy within a given collection or stream of data. For example, a major event, such as a natural disaster, may lead to hundreds or even thousands of highly redundant related news articles being written.

LSI provides a simple mechanism for dealing with such a situation. The basic concept is shown in Fig. 3.

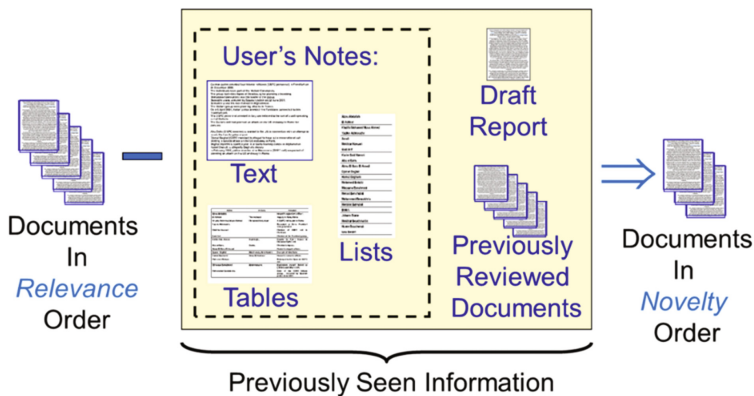


Fig. 3. Technique for detecting novel information

In this application, some retrieval mechanism is used to initially identify relevant documents. LSI vectors for those documents are compared to LSI vectors representing the user's existing knowledge. An estimate of this knowledge can be derived from artifacts such as previously reviewed documents, notes the user has taken, etc. The *relevant* documents are then re-ranked so that the ones containing the most *novel* information (i.e., having the lowest cosine scores with respect to the current user knowledge representation) are placed at the top of the data presented to the user.

### 4.4 Item-Item Relationship Discovery

Comparisons of lists of entities or concepts can be of particular value in scenario discovery. In one example, an LSI space was created from 158,492 English-language news articles [41]. The text was processed using an entity extractor to identify PEOPLE, ORGANIZATIONS, and LOCATIONS (including facilities). The use case was to identify associations between *terrorist groups* and *things that they might attack*. The entity extractor identified 170,479 named entities of type ORGANIZATION in the

collection. The LSI representation vectors for these ORGANIZATIONS were compared to a vector representing the concept of *terrorism*. The top 350 items from that comparison all corresponded to names of terrorist groups. This set was used as the list of terrorist groups for the analysis. All 37,706 of the identified LOCATIONS were considered as potential targets.

A 350 by 37,706 matrix of terrorist groups vs. potential targets was formed, with each entry being the cosine between the LSI representation vectors for the respective terrorist group and potential target. Table 2 shows some selected rows and columns from this matrix.

**Table 2.** Matrix of terrorist groups vs. potential targets

	Athens Airport	NATO HQ	Strasbourg Cathedral	Trafalgar Square	The Vatican
Abu Sayyaf	-.0254	-.0235	-.1235	-.0862	.0805
al Aqsa	-.0152	-.0572	-.0710	-.0120	-.0222
GSPC	-.1556	-.0043	.5087	.1677	-.1343
Hamas	.0071	.0462	-.0019	.0152	.0065
Hizballah	-.0513	.0028	-.0508	-.1056	.0458
Islamic jihad	.0138	.0300	-.0439	.0086	-.0133
PFLP	.0580	-.0190	.0101	.0931	.0365

The great majority of the entries in the matrix showed a very low degree of relatedness, with typical cosine values in the range of  $-.1$  to  $+1$ . The cosine value of  $.5087$  between the vector corresponding to the *GSPC*<sup>3</sup> and that for *Strasbourg Cathedral* was larger than that of any of the other 13 million matrix entries. This turned out to be a correct association. In the time frame of these articles, the Frankfurt cell of the *GSPC* indeed had planned an attack on the cathedral in Strasbourg. Of particular importance, in this document collection, no articles contained both the term *GSPC* and the term *Strasbourg Cathedral* (or any synonyms for these entities). Two articles did associate the names of specific individuals with *Strasbourg Cathedral*. Relationships among names in other articles, in aggregate, associated those individuals with the *GSPC*. The high cosine value here thus is based completely on LSI holistic analysis of *n*-th-order associations. This was a completely unexpected result. It is a good example of the first type of discovery cited in Sect. 4 – where the user initially does not have the necessary information to create an effective query for conventional tools such as Boolean search.

This is a fully general technique. It could be applied to compare any two lists of terms, entities, or even concepts. It could be implemented in a manner similar to that discussed in the following section in order to generate alerts of developing situations of potential interest.

<sup>3</sup> Subsequent to the timeframe covered by the news articles used in the experiment, the Salafist Group for Preaching and Combat (*GSPC*), changed its' name to Al Qaida in the Islamic Maghreb (*AQIM*).

### 4.5 Term-Concept Relationship Discovery

LSI demonstrably is capable of modeling relations through fifth order<sup>4</sup> [42]. Integration of such relations across a large document collection provides an ability to identify even quite subtle relationships between objects. This capability can be used effectively in automatically identifying entities of interest for a user. The workflow for an application of this type is shown in Fig. 4.

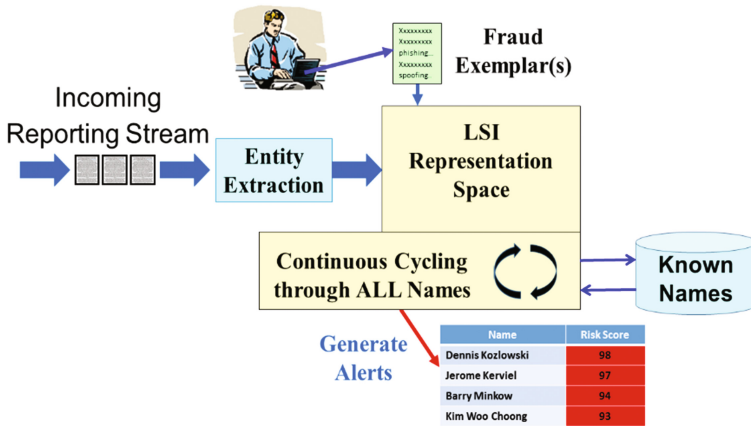


Fig. 4. Automatic identification of entities of interest

The objective in this example is to identify individuals who may have an association with fraud. The LSI space is continuously being updated with data from newly arriving documents. The system cycles through all named entities identified as PERSONS and compares the names to a semantic representation of the concept of fraud within the LSI space. Those names having an association with fraud that is above a threshold are identified. Identified names that are not already known are then presented to the user in the form of an alert. The key aspect of this workflow is that it examines *all* of the names in the collection. It thus can identify individuals of interest for whom the user had no previous reason to be suspicious (and thus would not have known to incorporate their name in a query).

### 4.6 Discovery Incorporating Spatiotemporal Data

Recently, work has been done on incorporating geospatial and temporal information into LSI-based analysis systems. Figure 5 shows the results of a query against a state-of-the-art LSI-based analysis system that integrates spatial, temporal, and conceptual data. The LSI space used to produce this display was based on a corpus of approximately 300,000 documents. The query dealt with social unrest.

<sup>4</sup> For example, Person A – Person B – Organization C - Telephone Number D – Person E – fraud.





Fig. 5. Integration of spatial, temporal, and conceptual information

What is unique about this application is that the displayed points are determined *conceptually*. A city may be highlighted even though it is not mentioned in *any* document that *directly* discusses the concept of interest (similar to the situation discussed in Sect. 4.4). Through exploitation of the conceptual generalization and *n*th-order analysis capabilities of LSI, this application can cause a location to be highlighted based on quite subtle information. This can constitute a powerful discovery capability.

## 5 Conclusion

As shown in the examples presented here, LSI is capable of providing a wide range of support to users in carrying out discovery activities. Overall, use of LSI in discovery applications is increasing in scale and sophistication.

There are a number of promising areas for future work. One interesting question is to compare how well LSI performs in discovery applications in comparison to other techniques. This is a gap in the existing literature.

Investigation of LSI-based discovery involving other data types also appears to be quite promising. Most of the LSI-based discovery applications to date have involved text in a single language. Application of the cross-lingual capabilities of LSI should enable exploitation of information resources in one language in order to facilitate discovery in text collections in other languages. More generally, as noted in Sect. 2,

LSI can be used for analysis of other types of data, including audio, images, signals, and video. Investigation of the use of LSI for discovery within collections of those types of data is likely to be particularly rewarding.

## References

1. Sadeh, T.: From search to discovery. In: World Library and Information Congress, Singapore (2013)
2. Bellegarda, J.: Latent semantic mapping. *IEEE Sig. Process. Mag.* **22**(5), 70–80 (2005)
3. Bradford, R.: Techniques for processing LSI queries incorporating phrases. In: 6th International Conference, IC3 K. CCIS, Rome, vol. 553, pp. 99–117. Springer (2014)
4. Furnas, G., et al.: Information retrieval using a singular value decomposition model of latent semantic structure. In: 11th SIGIR, Grenoble, France, pp. 465–480 (1988)
5. Bradford, R.: Comparability of LSI and human judgment in text analysis tasks. In: Applied Computing Conference, Athens, Greece, pp. 359–366 (2009)
6. Michel, K.: Personal communication, 14 April 2017
7. Oard, W., Webber, W.: Information retrieval for e-discovery. *Found. Trends Inf. Retrieval* **7**(2–3), 99–237 (2013)
8. McArthur, R., Bruza, P.: Discovery of implicit and explicit connections between people using email utterance. In: 8th European Conference on CSCW, pp. 21–40 (2003)
9. Skillicorn, D.: Detecting anomalies in graphs. Technical report # 2007-529, Queen's University, Ontario, Canada (2007)
10. Fortuna, B., Mladenich, D., Grobelnik, M.: Semi-automatic construction of topic ontologies. In: Semantics, Web and Mining. LNCS, vol. 4289, pp. 121–131. Springer, Heidelberg (2006)
11. Louwse, M., Zwaan, R.: Language encodes geographical information. *Cogn. Sci.* **33**, 51–73 (2009)
12. Lia, W., Goodchild, M., Raskinc, R.: Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *Int. J. Digital Earth* **7**(1), 17–37 (2014)
13. Fu, K., Cagan, J., Kotovsky, K.: A methodology for discovering structure in design data-bases. In: International Conference on Engineering Design, Denmark, vol. 6 (2011)
14. Vockner, B., Richter, A., Mittlböck, M.: From geoportals to geographic knowledge portals. *Int. J. Geo-Inf.* **2**(2), 256–275 (2013)
15. de Boer, R., Vliet, H.: Architectural knowledge discovery with latent semantic analysis: constructing a reading guide for software product audits. *J. Syst. Softw.* **81**(9), 1456–1469 (2008)
16. Kesorn, K.: Multi-modal multi-semantic image retrieval, Ph.D. thesis, University of London (2010)
17. Chen, X., et al.: A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval. In: 7th IEEE ISM, Taiwan (2005)
18. Jassez, J.-L., et al.: Signature based intrusion detection using latent semantic analysis. In: IJCNN, Hong Kong, pp. 1068–1074 (2008)
19. Pramanick, S., Rajagopalan, S., van den Berg, E.: Mitigating the insider threat with high-dimensional anomaly detection, AFRL-IF-RS-TR-2004-338, Final report (2004)
20. Zhu, W., Chen, C.: Storylines: visual exploration and analysis in latent semantic spaces. *Comput. Graph.* **31**(3), 338–349 (2007)
21. Freitas, A., Curry, E., Handschuh, S.: Towards a distributional semantic web stack. In: 10th International Workshop on Uncertainty Reasoning for the Semantic Web, pp. 49–52 (2014)

22. Ma, J., Zhang, Y., He, J.: Web services discovery based on latent semantic approach. In: IEEE International Conference on Web Services, Beijing, pp. 740–747 (2008)
23. Shahriar, H., Haddad, H.: Object injection vulnerability discovery based on latent semantic indexing. In: 31st Annual ACM SAC, Pisa, Italy, pp. 801–807 (2016)
24. Bhatia, L., Cao, K.: Intelligent polar infrastructure: enabling semantic search in geospatial metadata catalogue to support polar data discovery. *Earth Sci. Inform.* **8**(1), 111–123 (2015)
25. Hashimoto, T., Kuboyama, T., Chakraborty, B.: Temporal awareness of changes in afflicted people’s needs after the East Japan Great Earthquake. In: IEEE TENCON, pp. 1–6 (2013)
26. Speer, R., Havasi, C., Lieberman, H.: Analogy space: reducing the dimensionality of common sense knowledge. In: 23rd National Conference on Artificial Intelligence, pp. 548–553 (2008)
27. Keila, P., Skillicorn, D.: Detecting unusual and deceptive communication in email. Technical Report # 2005-498, Queen’s University, Ontario, Canada (2005)
28. Rossi, R.: Latent semantic analysis of the languages of life. In: 4th ISICA. CCIS, Huangshi, China. Springer, vol. 51, pp. 128–137 (2009)
29. Homayouni, R.: Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics* **21**(1), 104–115 (2005)
30. Gong, L., Yang, R., Yan, Q., Sun, X.: Prioritization of disease susceptibility genes using LSM/SVD. *IEEE Trans. Biomed. Eng.* **60**(12), 3410–3417 (2013)
31. Kim, H., Park, H.: Extracting unrecognized gene relationships from the biomedical literature via matrix factorizations using a priori knowledge of gene relationships. In: 1st International Workshop on Text Mining in Bioinformatics, Virginia, pp. 60–67 (2006)
32. Fukushima, A.: SVD-based anatomy of gene expressions for correlation analysis in Arabidopsis thaliana. *DNA Res.* **15**(6), 367–374 (2008)
33. Vanteru, B., Shaik, J., Teasin, M.: Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genom.* **9**(Suppl 1), S10 (2008). *BIOCOMP 2007*
34. Roy, S., et al.: Latent semantic indexing of PubMed abstracts for identification of transcription factor candidates from microarray derived gene sets. *BMC Bioinform.* **12** (Suppl 10), S19 (2011)
35. Xu, L., et al.: Functional cohesion of gene sets determined by latent semantic indexing of PubMed abstracts. *PLoS ONE* **6**(4), e18851 (2011)
36. Wei, L., et al.: Inferring gene regulatory mechanisms from microarray data using latent semantic indexing of MEDLINE abstracts: the role of Rel in Type-I interferon signaling. *FASEB J.* **20**, A929 (2006)
37. Doong, S., Hong, S-F.: Protein-protein interaction document mining. *Advances in Intelligent Systems Research* (2006)
38. Dos Santos, E., et al.: A semantic-based similarity measure for human druggable target proteins. In: The Fifth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO2013), Lisbon, Portugal, March 24–29 (2013)
39. Bradford, R.: Efficient discovery of new information in large text databases. In: *Intelligence and Security Informatics*. LNCS, vol. 3495, pp. 374–380. Springer (2005)
40. Bradford, R.: Use of latent semantic indexing to identify name variants in large data collections. In: *IEEE Intelligence and Security Informatics*, pp. 27–32 (2013)
41. Bradford, R.: Relationship discovery in large text collections using latent semantic indexing. In: *SIAM Data Mining Conference, Workshop on Link Analysis, Counterterrorism and Security*, Bethesda, Maryland (2006)
42. Kontostathis, A., Pottenger, W.: Mathematical view of latent semantic indexing: tracing term co-occurrences. Technical Report LU-CSE-02-006, Department of Computer Science and Engineering, Lehigh University (2002)

# Similarity Measures for Music Information Retrieval

Michele Della Ventura<sup>(✉)</sup>

Department of Technology, Music Academy “Studio Musica”, Treviso, Italy  
dellaventura.michele@tin.it

**Abstract.** This article is aimed at presenting a method for the assessment of the similarity between two data strings representing the musical text analyzed on a symbolic level (music notes), in order to cluster and classify musical pieces with particular reference to the files stored according to the MIDI standards. This method is based on the quantification of each string by calculating its entropy. It is an empirical methodology that provides results expressed in numbers that may be analyzed. This enables a comparison between two strings of different length highlighting their potential identity (sameness), similarity and homology. The representation by means of an isometric diagram accommodates a better interpretation of the results. The effectiveness of the proposed melodic representations and algorithms is tested against a series of melodic examples.

**Keywords:** Entropy · Homology · Information retrieval · Similarity

## 1 Introduction

One of the main purposes in the Artificial Intelligence field is to support the individuals’ decision-making process. To that end, the ability to obtain information relevant to the task at hand is of utmost importance.

The ever growing number of digital information present on the web made the search engine an indispensable tool on which all the users depend when it comes to identifying the desired piece of information. One of the issues is that the search engine must tackle in order to meet the requirements of the users is to assess whether a textual piece of information is relevant in relation to the information provided by the user.

This issue of the research drew the attention of many scholars and is still far from being completely solved. In effect, the web is a constantly evolving tool that keeps offering new opportunities in various fields and among them, the music field. The connection between music and the digital field has occurred (and is developing) both on the audio level, with the possibility to share a music file to listen to, and on the textual level, with the possibility to share a music score considered on a symbolic level, i.e. the notes. In the latter case methods have been developed to retrieve the information, all based on the concept of similarity between the character string indicated by the user and the string taken into consideration by the search engine (within the various music scores existing in a database).

The problem of Symbolic Melodic Similarity, where a retrieval system is expected to get back a list of musical pieces considered similar to another one, has been studied

from different standpoints. Some techniques are based on geometric representations of music [1–4], others depend on classic grammar rules representations to determine similarities [5, 6], and others use editing distances and alignment algorithms [7–9].

Cambouropoulos has a different approach that focuses on similarity. He has proposed [10, 11] a computational model where the notions of categorization, similarity, and the representation of entities/properties are strongly connected. It is not simply the case that one starts with a precise description of entities and properties, then finds similarities between them and, finally, gathers the most similar ones together into categories.

Cambouropoulos [12] also shows the necessity for researchers to state clearly what view and what definitions of similarity/categorization they support in order to avoid superfluous conflict and disorientation. The distinction between similarity and identity (identicalness) of two strings is very rigorous: “*Similarity is very often defined as partial identity, that is, two entities are similar if they share some properties, but not necessarily all*”.

This article presents a method for the analysis of the similarity that takes in considerations the analyzed strings as separate entities which, however, belong to the same musical text. This enables a comparison based on the concept of “entropy”, by means of a quantification of the information carried along by every single sound: the single sound is the smallest information unit. The comparison of the information values of every single sound enables us to identify which parts of the strings are identical or similar and, finally, their potential homology.

This paper is organized as follows.

Section 2 describes the concepts of identity, similarity and homology. Section 3 describes the information theory. Section 4 describes the analysis of the musical message. Section 5 shows some experimental tests that illustrate the effectiveness of the proposed method. Finally, conclusions are drawn in Sect. 6.

## 2 Identity, Similarity and Melodic Analogy

A musical fragment is represented by a succession of sounds (melody) of different pitches and duration. Melody and rhythm are two fundamental elements for musical architecture (structuring), two almost inseparable elements: a melody develops on the rhythm and without it does not exist [13].

Information Retrieval is based on the comparison between two musical fragments:

1. A model fragment, which is represented by the notes intentionally inserted by the user;
2. A sample fragment, which is represented by the notes extrapolated from a digital score by the search engine.

On the basis of these assumptions, by comparing two musical fragments the following possibilities may be identified:

1. Identity (Fig. 1a): the distance expressed in semitones (interval) that separates each sound of the model fragment (Fig. 1 the first stave) is the same as the distance separating each sound of the sample fragment (Fig. 1 second stave); the number will be positive in case of an ascending interval and negative in case of a descending interval;

2. Similarity (Fig. 1b): the pace (a = ascending or d = descending) separating each sound of the model fragment is equal to the pace separating each sound of the sample fragment, regardless of the values of their intervals;
3. Homology (Fig. 1c): the two musical fragments have similar traits but morphologically they may be different because they are modified so as to be functional.



**Fig. 1.** Comparison of two: identical melodies (a), similar melodies (b) and homologous melodies (c)

When comparing two musical fragments it is, therefore, not important whether the sound sequences are different: the sound is the fundamental element to start the analysis with, considering its (ascending or descending) movement and the function it performs. Thus it is possible to consider the set of sounds of two fragments as a single unit of analysis, i.e. a score, and to analyze the various internal functions of the latter by quantifying its information.

### 3 Information Theory

The analysis based on information theory sees music as a linear process supported by a syntax of its own [14]. However we are approaching to a syntax formulated not on the basis of musical grammar rules, but on the basis of the probabilities of occurrence of every single element of the musical message in relation their preceding element [15]. From the definition of “message” being a chain of discontinuous “units of meaning” follows that the musical “units of meaning” correspond to the minimal events of a composition: usually isolated notes, chords...

Any event of such a chain requests a prevision about the event that will follow: there is information transmission when the prevision is unexpected; there is no information transmission when it is confirmed. Moreover, the fact that the events of a composition are organized modularly brings forward the possibility to calculate using a formula, or to express with an “index” the total amount of information transmitted by a certain musical segment [16]. In a communication that occurs by means of a given alphabet of symbols, information is associated to every single transmitted symbol. Hence, the information may be defined as the reduction of uncertainty that might a priori have existed in the transmitted symbol [15, 17].

The wider the message range that the source can transmit is (and the greater the uncertainty of the receiver with respect to the possible message), the larger the quantity of transmitted information and along with it its measure: the entropy.

In the information theory, entropy is a positive value as opposed to its original negative counterpart in physics. Mathematically, the measure of the content of a piece of information ( $I$ ) is obtained with Shannon's formula:

$$I = \log_2 \frac{p'}{p} \quad (1)$$

where  $p$  is the probability of the message to be transmitted,  $p'$  corresponds to the probability for the content of the information expected by the latter to be satisfied after the transmission of the message. For every symbol (of a message) that we transmit we have a certain quantity of information associated to that symbol.

In most practical applications of information theory a choice must be made among the messages of a set, every single one having its own probability of being transmitted.

Shannon gave a definition of the entropy of such a set, identifying it with the information content that the choice of one of the messages will transmit. If every single message has the probability  $p_i$  of being transmitted, the entropy is obtained as the sum of all the set of functions  $p_i \log_2 p_i$ , every single one related to a message, that is:

$$H(X) = E[I(x_i)] = \sum_{i=1}^n I(x_i) \cdot P(x_i) = \sum_{i=1}^n P(x_i) \cdot \log_2 \frac{1}{P(x_i)} \quad (2)$$

The term entropy, borrowed from thermodynamics, designates therefore the average information content of a message. In the light of what has been said so far, the musical message may be defined as a sequence of signals organized according to a code.

## 4 Analysis of the Musical Message

To compare various segments among them, in order to determine which is more important, each entropy is calculated: the less the entropy value, the greater the information carried by the sound [18].

In order to calculate the entropy it is necessary to take into consideration a specific alphabet: the alphabet is language – specific [19–21] and, as it may be immediately deduced from the formula (based on the probability of certain symbols rather than other symbols to be transmitted) it demonstrates to be associated to language.

For the **melodic analysis** the various melodic intervals were classified as symbols of the alphabet [18].

The classification of an interval consists in the denomination (generic indication) and in the qualification (specific indication).

The denomination corresponds to the number of degrees that the interval includes, calculated from the lowest one to the highest one; it may be of a 2<sup>nd</sup>, a 3<sup>rd</sup>, 4<sup>th</sup>, 5<sup>th</sup>, and so on.; the qualification is deduced from the number of tones and semi-tones that the interval contains; it may be: perfect, major, minor, augmented, diminished, more than augmented, more than diminished, exceeding, deficient.





## 5 Obtained Results

The model of analysis described in this article was verified by realizing an algorithm the structure of which takes into consideration each and every single aspect described above. The results of the analysis are indicated by means of isometric graphics that allow an immediate visualization for interpretative purposes. The algorithm does not provide any limitation with respect to the dimensions of the table representing the alphabet and the matrix of transitions that will be automatically dimensioned in every single analysis on the basis of the characteristics of the analyzed musical segment. This allows conferring generality to the algorithm and specificity to every single analysis (Strength).

The initial tests were carried out on a set of musical segments of various lengths, specifically selected, in order to verify the “validity” of the analysis. Subsequently, entire scores were taken into consideration and subjected to segmentation by the algorithm in order to identify and classify the musical segments, by comparison with a model segment entered by the user.

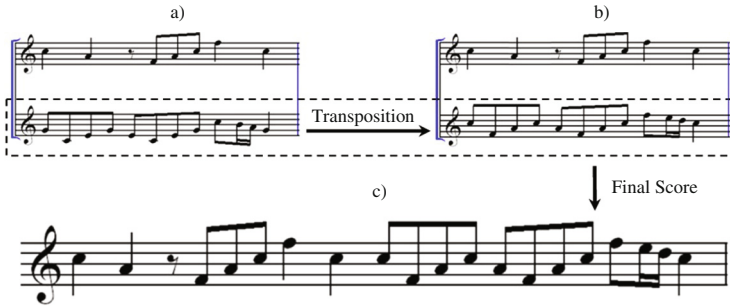
The algorithm carries out the analysis steps considering two segments, a model and a sample, as if they were one single score: the various sounds are considered one after another (Fig. 2c) in order to determine the alphabet, the transition table and therefore quantify the information value of each sound by calculating the entropy. The objectives are:

1. To determine the distinct brackets of distinctive values of the information content, by determining a minimum and a maximum value, which may be deduced from the information values of every single sound;
2. Verify, for each sound of the two segments being compared, the bracket within which they identify themselves: the value of the first sound of the model segment will be compared to the value of the first sound of the sample segment, and so on. If the information value of the two sounds falls within the same bracket, they share equivalence/similarity.

This procedure is due to the fact that, according to the information theory exposed above, the information value depends on the alphabet of the score which determines the transition probabilities among the states of the system (*conditional probability*): the probability, in our case, for a sound to resolve to another sound (giving birth to an interval). The distance between various sounds and their ascending or descending movement somehow represents the DNA of the composition.

Some illustrative results are presented below in relation to the comparison between two different segments so as to explain, by means of graphic representation, how the data is interpreted.

Two musical segments are represented in Fig. 2c: a model segment (the first stave) and a sample segment (the second stave). In case the two segments (model and sample) are in different tonalities, the algorithm first operates a transposition of the sounds of the sample segment (Fig. 2b) taking them to the same pitch as the ones of the model segment.



**Fig. 2.** Score representation

Once the various tables (alphabet and transitions) have been defined, the information value of every single sound is calculated (Table 3) and then the information brackets are identified, each of them representing a certain range of values.

**Table 3.** Information value of every single sound

Model segment	Sample segment	
0.530737271	0.530737271	
0.464385619	0.528320834	
0.389975	0.442179356	
0.442179356	0.523882466	
0.523882466	0.464385619	
0	0.389975	
0.401050703	0.442179356	
	0.523882466	
	0	
	0	
	0	
	0.401050703	
Bracket 1	Bracket 2	Bracket 3
0 – 0,2	0.2 – 0,4	0.4 – 0,6

Figure 3 shows two segments compared and the related diagram. The upper part of the diagram represents through colors the various brackets of information of the individual sounds of the model segment, while the lower part of the diagram shows the various brackets if information of the individual sounds of the sample segment. The color representing a certain bracket will have a darker or clearer tone based on whether the information value is higher or lower within the same identification range of the same bracket. If two sounds have the same information value the diagram will contain a column having only the color of the corresponding information bracket. If, instead, the two values are different, the color within the column changes by fading out, passing from the color of the preceding value to the color of the current value. The bigger the color difference, the smaller the equivalence or the similarity between the segments (Fig. 4).

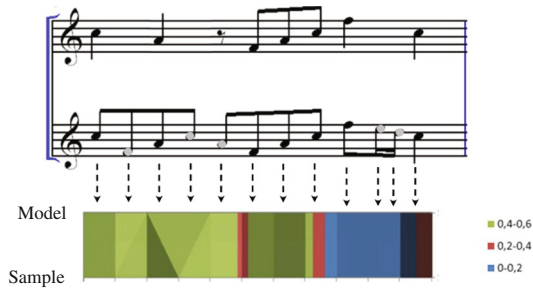
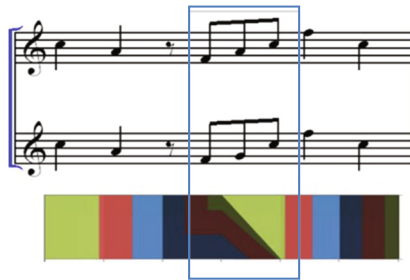


Fig. 3. Isometric diagram



Model segment		Sample segment
0.52388247		0.52388247
0.389975		0.389975
0		0
0		0.52832083
0.43082708		0.43082708
0		0
0.52832083		0.52832083
Bracket 1	Bracket 2	Bracket 3
0 - 0,2	0,2 - 0,4	0,4 - 0,6

Fig. 4. Determination of the dissimilarity of the two segments.

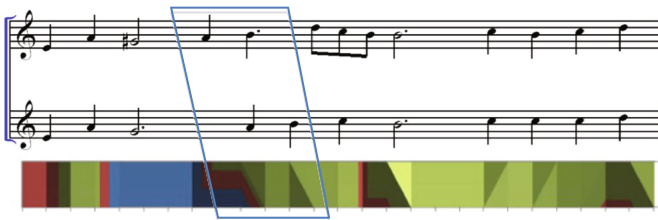


Fig. 5. Determination of the dissimilarity of the two segments.

Figure 5 represents two musical segments where the first one is different from the second both in terms of rhythm and because of the presence of melodic figurations. However, the two segments draw attention to a clear similarity between them.

The color column of every single sound belong to the same range even if with different values (see the undertones of green), except for the highlighted point where the sounds of the second segment are equal to the ones of the first but out of phase in time (delayed): the color though fades towards the color of the sound of the first segment.

## 6 Discussion and Conclusions

In recent years there has been a continuous growth of the number of databases used for storage and sharing of musical information (in a digital format), be them on-line or not: information that is not only textual, but also symbolic, such as for instance the scores and their representation through the MIDI protocol recently reinstated by the web. Simultaneously to the development of the databases there has been a development of algorithms for Information Retrieval, which, however, gave the right answer to the retrieval of musical information. In general this retrieval operation is based on the similarity concept.

This study took into consideration the possibility to compare two musical segments, analyzed on a symbolic level, by measuring the information content of each sound. The proposed approach had the objective of representing the relationships among the individual musical segments by means of an isometric diagram, built on the basis of the entropy concept. This allowed the determination of the potential equivalence, similarity and homology of the analyzed segments.

The approach has been tested with a collection of musical segments. Besides, qualitative analysis has been realized relative to the relationships between the graph structure and the melodic content of the musical segment. Results are interesting in terms of average precision of the retrieval results and in terms of musicological significance.

This method of analysis might be extended to the process of segmentation of a musical composition and, therefore, represent an important means of support for the teacher and for the student of compositive disciplines. Finally, it could be extended to the studies related to Automated Speech Recognition (ASR), in particular, for the dictation, not of words, but of musical phrases.

## References

1. Aloupis, G., et al.: Algorithms for computing geometric measures of melodic similarity. *Comput. Music J.* **30**(3), 67–76 (2006)
2. Laitinen, M., Lemström, K.: Geometric algorithms for melodic similarity. In: Proceedings of the Annual Music Information Retrieval Evaluation Exchange (2010). [music-ir.org/mirex/abstracts/2010/LL1.pdf](http://music-ir.org/mirex/abstracts/2010/LL1.pdf). Accessed Feb 2015
3. Lemström, K.: Towards more robust geometric content-based music retrieval. In: Proceedings of the Conference of the International Society for Music Information Retrieval, pp. 577–582 (2010)
4. Urbano, J.: A geometric model supported with hybrid sequence alignment. In: Proceedings of the Annual Music Information Retrieval Evaluation Exchange (2013). [music-ir.org/mirex/abstracts/2013/JU1.pdf](http://music-ir.org/mirex/abstracts/2013/JU1.pdf). Accessed Feb 2013

5. Yazawa, S., et al.: Melodic similarity based on extension implication-realization model. In: MIREX Symbolic Melodic Similarity Results (2013). [music-ir.org/mirex/abstracts/2013/YHKH1.pdf](http://music-ir.org/mirex/abstracts/2013/YHKH1.pdf). Accessed Feb 2013
6. Pancini, M.: Problematiche e modelli formali per la segmentazione automatica/interattiva di partiture musicali simboliche, Rapporto Tecnico CNR-PFBC-MUS-TR (2000)
7. Gibbs, A.J., Mcintyre, G.A.: The diagram. A method for comparing sequences. *Eur. J. Biochem.* **16**, 1–11 (1970)
8. Vempala, N.N., Russo, F.A.: An empirically derived measure of melodic similarity. *J. New Music Res.* **44**(4), 391–404 (2015)
9. Urbano, J., Lloréns, J., Morato, J., Sánchez-Cuadrado, S.: Using the shape of music to compute the similarity between symbolic musical pieces. In: International Symposium on Computer Music Modeling and Retrieval, pp. 385–396 (2010)
10. Cambouropoulos, E.: Towards a general computational theory of musical structure. Ph.D. Thesis, Faculty of Music and Department of Artificial Intelligence, University of Edinburgh (1998)
11. Cambouropoulos, E.: Melodic cue abstraction, similarity and category formation: a formal model. *Music Percept.* **18**(3), 347–370 (2001)
12. Cambouropoulos, E.: How Similar is Similar? *Music Musicae Scientiae, Discussion Forum 4B*, pp. 7–24 (2009)
13. Orff, C.: *Schulwerk, elementare Musik*. Hans Schneider, Tutzing (1976)
14. Weaver, W., Shannon, C.: *The Mathematical Theory of Information*. Illinois Press, Urbana (1964)
15. Angeleri, E.: *Information, Meaning and Universalit*. UTET, Turin (2000)
16. Moles, A.: *Teorie de l'information et Perception esthetique*. Flammarion Editeur, Paris (1958)
17. Lerdhal, F., Jackendoff, R.: *A Grammatical Parallel Between Music and Language*. Plenum Press, New York (1982)
18. Della Ventura, M.: Analysis of algorithms' implementation for melodic operators in symbolical textual segmentation and connected evaluation of musical entropy. In: *Proceedings of 1st Models and Methods in Applied Sciences, Drobeta Turnu Severin*, pp. 66–73 (2011)
19. Monelle, R.: *Linguistics and Semiotics in Music*. Harwood Academic Publisher, Chur (1982)
20. Della Ventura, M.: Rhythm analysis of the “sonorous continuum” and conjoint evaluation of the musical entropy. In: *Proceedings of Latest Advances in Acoustics and Music, Iasi*, pp. 16–21 (2012)
21. Nattiez, J.J.: *Fondements d'une sémiologie de la musique*. Union Générale d'Éditions, Paris (1975)

# An Early-Biologisation Process to Improve the Acceptance of Biomimetics in Organizations

Nguyen-Truong Le<sup>(✉)</sup>, Joachim Warschat, and Tobias Farrenkopf

Fraunhofer IAO, Stuttgart, Germany  
truong.le@iao.fraunhofer.de

**Abstract.** Biologisation of technology has already become an innovation driver for new materials and implants in medical applications. However, the potential of biologisation of technology is much greater when taking all the current research fields into consideration: e.g. the process industries using microorganisms for production of new substances or pharmaceuticals, the IT developing evolutionary algorithms for autonomous systems supporting decision making of human being, or various engineering disciplines adapting principles from nature to optimize both resource consumption and performance.

How can technical and scientific disciplines systematically learn from nature and capture value from biologisation in future? In this paper an early-biologisation approach will be introduced. When combining with the biomimetic engineering process according to ISO 18458:2015 the two approaches would improve the acceptance of biomimetics in organizations.

## 1 Introduction

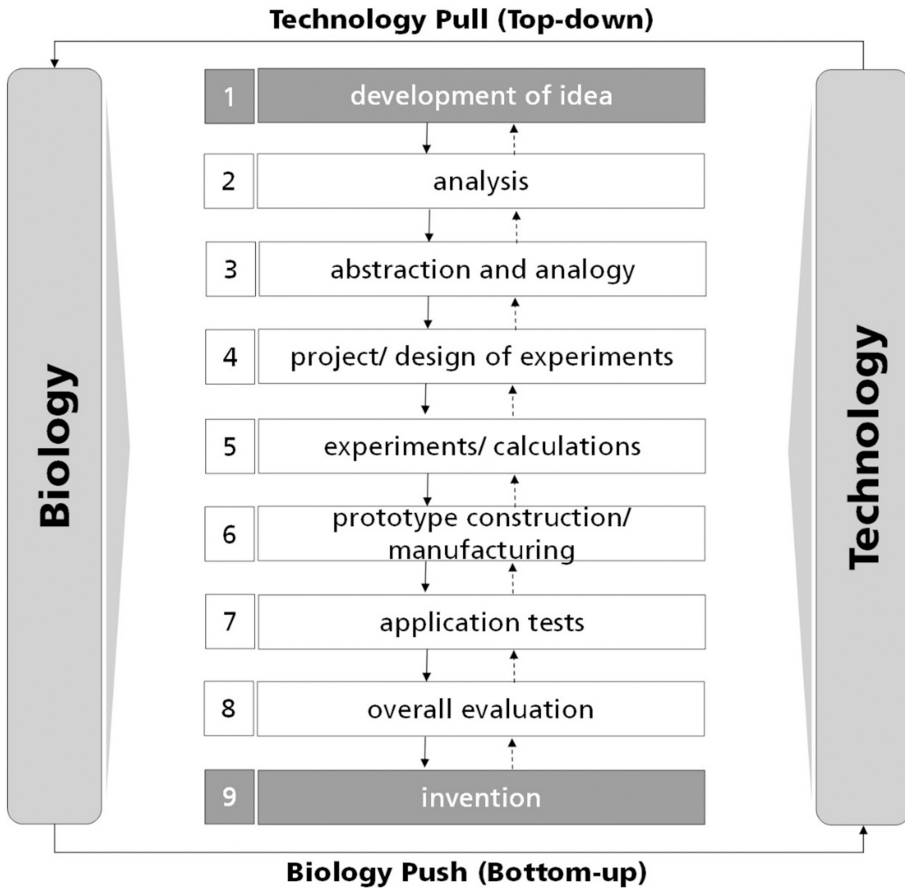
Nature has always been a main source of inspiration for mankind. Through the study of animals or plants important inventions like aircraft<sup>1</sup>, reinforced concrete<sup>2</sup> were born. The term “biomimetics” stands for the creative transfer of knowledge and ideas from biology to technology [1]. Nature still represents an inexhaustible source of ideas for dealing with various daily issues such as biocompatible implants, energy and resource saving in transportation, or object recognition algorithms. Current approaches of learning from nature often follow two typical patterns [2, 3]:

1. **Biology-push approach:** The biology-push approach starts with a discovery in biology, which is analyzed and technically implemented.
2. **Technology-pull approach:** The starting point of the technology pull approach is a problem to be solved from the side of technology for which solutions from nature are looked for.

---

<sup>1</sup> Leonardo da Vinci (1452–1519) studied the anatomy and flight of birds when developing concepts for flying machines.

<sup>2</sup> Joseph Monier (1823–1906) was a gardener who found in the rigid network structure of the cactus plant *Opuntia* the inspiration for combining steel mesh with cement material to create reinforced concrete.



**Fig. 1.** The biomimetic engineering process ISO 18458:2015 [4]

ISO 18458:2015 suggests an ideal process for biomimetic engineering with nine steps. However in reality each step in the process could be very complex when knowledge is created and shared among biologists, engineers, and other disciplines. Several questions remained unanswered. E.g. in the real world it is still unclear how communication and knowledge barriers between different disciplines can be overcome. Another problem is the lack of acceptance within an organization for biomimetics. Technical staff members suggesting biomimetics in R&D projects could be regarded as exotic thinkers within their organization. Furthermore searching for biological problem-solving strategies is still an obstacle to engineers who are not trained in searching and reading biology literature. Last but not least, the lack of a systematic method to convert an identified biological role model into a technical solution is another challenge. In most cases, this step is done more intuitively and without methodological support. Biomimetics has already become a well-accepted scientific discipline. However in order to become an industry-accepted method the above mentioned obstacles need to be removed (Fig. 1).

In this paper the biomimetic engineering process according to ISO 18458:2015 will therefore be extended by implementing knowledge sharing mechanisms between people inside and outside of an organization. Two suggestions for improvement will be done:

- An early-biologisation approach to build-up a clear understanding why biomimetics needs to be employed by an organization to solve problems
- A knowledge sharing database for different disciplines when practicing learning from nature

The final aim is to enhance a shared understanding between stakeholders within an organization which really wants to build up a biologisation competency.

## 2 Biomimetic Engineering Process Should Start with an Early-Biologisation Process

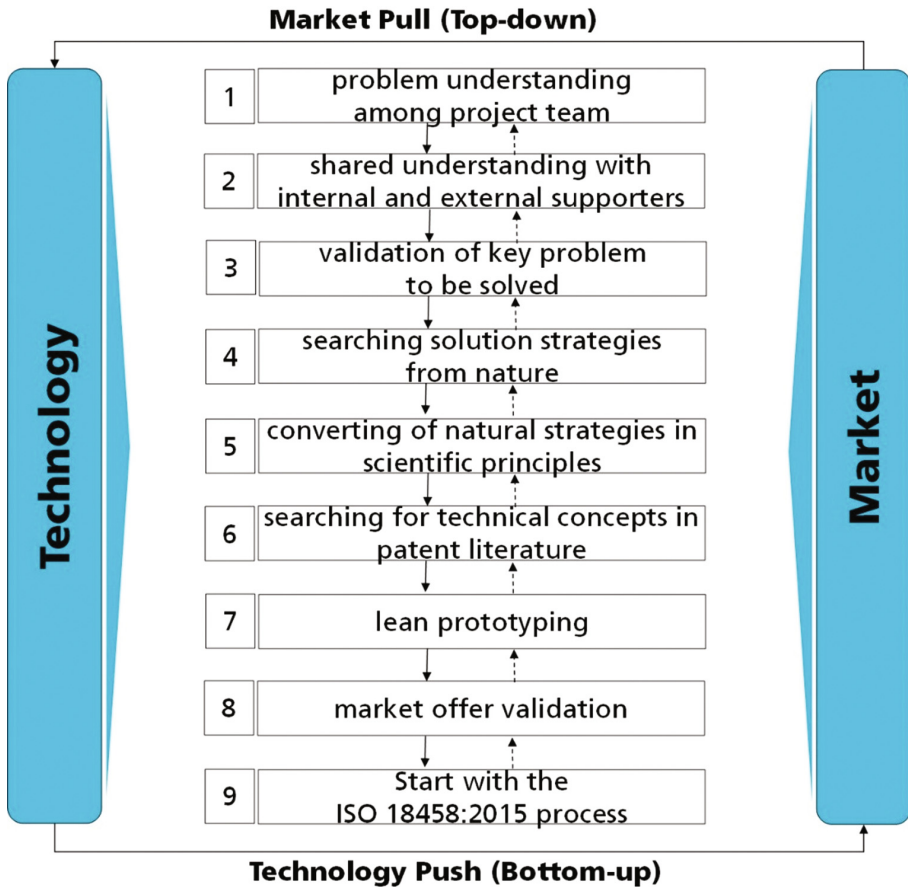
Early-biologisation is the building-up process of the competence “learning from nature” for an organisation prior to the biomimetic engineering process. The basic idea behind this is to create a shared understanding of problems to be solved by learning from nature. Otherwise the sophisticated process of biomimetic engineering process cannot gain support within an organization. Einstein has been often quoted “If I had an hour to solve a problem I’d spend 55 min thinking about the problem and 5 min thinking about solutions.” In this step an organization needs to invest time to check for the right problems to be solved. Secondly, within the process of biomimetic engineering team members need to enhance the radius of further supporters and co-creators inside and outside the organization. Thirdly, the value-added of biologisation of technology needs to be validated to gain insights into market potential. By the end of this early biologisation process an organization would be able to answer the question “why should we learn from nature to solve our problem?”. The early-biologisation process would deliver facts and insights before substantial investment decisions of R&D would be made (Fig. 2).

**Problem Understanding:** This step is about to understand the problem as much as possible. Information about: what exactly is given (dates, conditions, terms, etc.), what is searched, etc. is needed.

**Sharing Understanding with Internal and External Supporters:** It has been shown that project champions are essential for the success of innovation projects. Project champions could also be from outside an organization.

**Validation of Key Problems to be Solved:** Often, it turns out that the alleged problem and the real problem are quite different. Therefore, the reach of a clear and unambiguous understanding of the problem is essential because it could help to prevent wasting time, e.g. due to frequent changes to the project goals. Different methods e.g. value proposition design can be used for validation of problems.





**Fig. 2.** Proposal of an early-biologisation process prior to ISO 18458:2015.

**Searching Solution Strategies from Nature:** Once the problem is understood solution idea will be looked for. Usually having a flash of thought or an “enlightenment” requires an intensive study of the problem. Here, the use of creative methods could facilitate and accelerate the search for solutions.

**Converting of Natural Strategies into Scientific Principles:** This step is crucial in order to bridge the communication gap between disciplines. By converting natural phenomena e.g. the adhesion of gecko feet into scientific principle like the Van der Waals forces a common understanding can be achieved by project team members.

**Searching for Technical Concepts in Patent Literature:** Scientific working principles e.g. the Van der Waals forces are mostly too abstract for communication among project team members and stakeholders. It is therefore beneficial to link those scientific principles to existing working examples in the technical domain. Patent literature would be a good choice due to the huge information base and the easy-to-search structure of bibliographic data.

**Lean Prototyping:** To support the communication with internal and external stakeholders preliminary prototypes based on the idea of “lean start-up method” are of great use [5]. The idea is to make use of simple means to build up an understanding about the solution idea. This can be realised as a mock-up, as a design study or as a design sketch.

**Market Offer Validation:** A brilliant idea is not useful if only some people believe in it. In this phase the prototypes will be presented to internal and external parties. The solution ideas need to collect evidences that they would address an urgent need [6].

After this final step the decision should be made whether a biomimetic project should be initiated employing the biomimetic engineering process according to ISO 18458:2015. At this point a clear understanding among stakeholders has been created.

### 3 Supporting the Search for Nature’s Solution Strategies with a Technology-Biology-Dictionary

#### 3.1 The Basic Concept of a Technology-Biology Dictionary

Several organizations have started to set-up databases about biological phenomena that might be of interest to engineers. Due to the huge effort to manually build up such databases both in time and human resources, those working groups are facing great difficulty in continuing with their project. Because of limitation on scope and depth of information, such databases have not the potential to cover every field of technology.

In order to support scientists and engineers having access to nature’s huge potential of problem solving strategies, the tool BIOPS<sup>®</sup> (BIOlogy inspired Problem Solving) has been developed at Fraunhofer IAO. Figure 3 illustrates the key elements of BIOPS<sup>®</sup>. The dictionary combines a manually compiled dictionary of more than 2000 solution strategies from nature with an automatically generated technology thesaurus of more than 9 million entries relating to technical problems. This combination enables users to find appropriate solution strategies for various technical fields.

The German demo version of BIOPS<sup>®</sup> is available online at [www.nature4innovation.com](http://www.nature4innovation.com). By using BIOPS<sup>®</sup> users can easily identify nature’s solution strategies, e.g. the tool is capable of matching a technical term like “icing” to related biological terms like “Siberian salamander”, “snow algae” or “Alaska beetle” [7].

#### 3.2 Automatic Generation of a Technology Thesaurus

A set of 2000 solution strategies from nature can be used for a much greater number of technical problems. E.g. a hydrophobic structure from nature can be used for an easy-to-clean surface or to solve an anti-icing problem. It is therefore beneficial to link similar technical terms by using a thesaurus. The manual creation process of a technology thesaurus would be of high effort. Therefore the automatic approach using the method co-occurrence analysis was chosen [8]. Following steps of data analysis were undertaken as shown in Fig. 4:

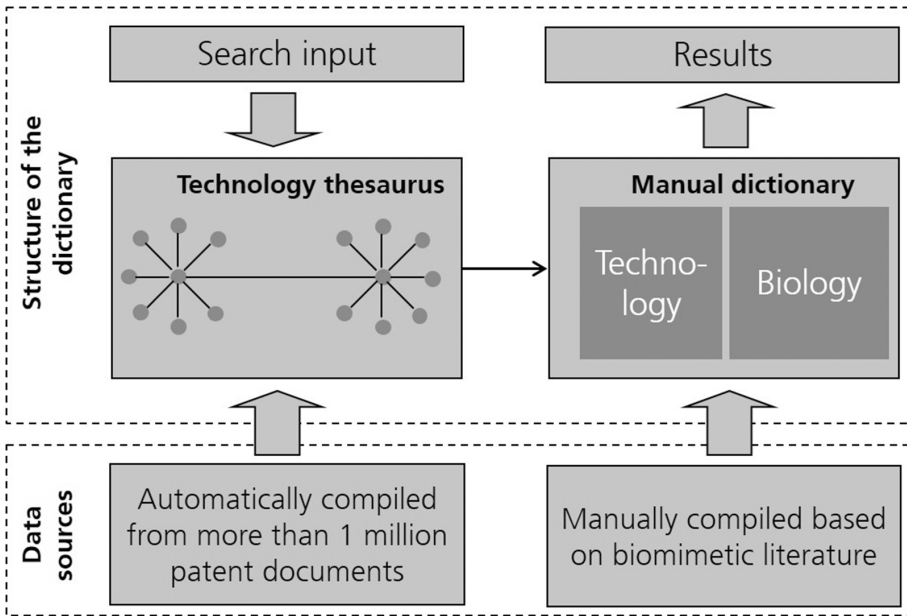


Fig. 3. Data structure of Technology-Biology-Dictionary BIOPS®.



Fig. 4. The automatic technology thesaurus generation process.

**Data Acquisition:** In the first step the data was acquired from more than 1 million patent documents. The greatest benefit of patent literature is that the verbs are mostly written in present tense. Therefore the algorithm for text analysis can be kept simple. Every single word of each sentence in the text corpus was reduced to its basic form.

*[The] [lotus] [effect] [refer] [to] [self-cleaning] [property]  
 [Barthlott] [and] [Ehler] [study] [the] [self-cleaning] [property]  
 [The] [lotus] [effect] [discovery] [open] [up] [new] [application]*

**Linguistic Preparation:** In the next step irrelevant words will be removed. Three methods has been employed to enhance the data quality:

- Creation of a black list with words that should be dropped
- Ignoring words of high term frequency (e.g. the, this, at, on etc.)
- Ignoring words of low term frequency

**Co-occurrence Analysis:** This method is used to analyse semantic proximity of words. In this work this method has been very useful to link similar technological terms with each other. In the example above the co-occurrences of the term [*self-cleaning*] are:

*[lotus], [effect], [refer], [property], [Barthlott], [Ehler]*

**N-gram Analysis:** The analysis of semantic relationship between words could be improved by using bi- or tri-grams instead of monograms. In this work the bigrams have been chosen. For the example “self-cleaning property” following bigrams has been extracted:

*[Lotus effect], [Barthlott Ehler], [new application]*

## 4 Evaluation

Since 2009 both the tool BIOPS and the early-biologisation approach have been used for 10 R&D projects at Fraunhofer IAO in cooperation with industrial partners. A project is regarded as successful when at least three from four criteria is fulfilled:

- A relevant problem of economic impact has been identified
- The solution is superior to existing solutions in terms of customer benefits (cost, quality, time)
- At least a solution has been identified which can be realised within short-term
- At least a solution has been identified which can be realised within long-term

**Table 1.** Success rate of past projects using the early-biologisation approach.

Industry	No. of successful projects	No. of failure
Mechanical engineering	3 projects	0
ICT	1 project	0
Sport equipment	1 project	0
Automobile	1 project	0
White goods	1 project	0
Chemical & food industry	2 projects	1
Success rate	<b>90%</b>	<b>10%</b>

Based on the experiences collected and summarized in Table 1 the early-biologisation approach has promising potential. In the one project considered as failure appropriate solutions could not be found by using the Technology-Biology dictionary. Therefore no solution could be realised within short-term and long-term.

In the project with the sport equipment manufacturer Fischer Sports GmbH a new generation of ski skin was developed based on biological role models e.g. snakeskin, sandfish and the mountain goat’s feet. Figure 5 provides a picture of the product PROFOIL as an outcome of the project with Fischer Sports GmbH. This project can be considered as a successful project due to three criteria:



**Fig. 5.** An output from a successful project using the early-biologisation approach. (Image provided by Fischer Sports GmbH to Fraunhofer IAO)

- Relevant problem: cross-country ski skins in the past requires intensive care due to snow adherence and icing
- Superior solution: the new solution approach is water repellent and easy-to-handle for users
- Solution is realised within short-term: the product can be now purchased

## 5 Implications

### 5.1 Implications for Further Research

The early-biologisation process represents a useful supplement to existing biomimetic engineering process. However there are still several research activities with regard to the search tool BIOPS<sup>®</sup> to be done in future:

- The BIOPS<sup>®</sup> tool is still a scientific search tool with simple user-interface. In future solution strategies from nature should be visualized using images or videos.
- The matching of relevant terms should be improved by using machine-learning techniques with information delivered by users.
- Last but not least latest advances in service-robotics should be explored to develop a brainstorming robot using the huge knowledge of BIOPS<sup>®</sup> that can be involved in creativity workshop with human beings. This would unleash new innovation potentials in future.

## 5.2 Implications for Practitioners

Better methods and tools to support the innovation process are only of high value when managers and engineers have sufficient knowledge about the pros and cons of using biomimetics in the industrial context. Not for every problem, biomimetics will give the appropriate answer. It is necessary to carry out an assessment for each company, whether biomimetics has any answer to the problems to be solved in their industrial context. In the next step, a pilot project related on a real problem to be solved should be organized. Based on the experiences collected in the pilot project further steps can be carried out to integrate biomimetics into the innovation process.

## References

1. Speck, O., Speck, T.: Bionische Innovationen, TEC21, 2002, 37 and 38, 22–25 (2002)
2. Speck, O., Speck, T.: Process sequences in biomimetic research. In: Brebbia, C.A. (ed.) Design and Nature IV, pp. 3–11. WIT Press, Southampton (2008)
3. Fratzl, P.: Biomimetic materials research – what can we really learn from nature’s structural materials? J. Roy. Soc. Interface **4**, 637–642 (2002)
4. International Organization for Standardization: Biomimetics – Terminology, concepts and methodology (2015). <https://www.iso.org/standard/62500.html>
5. Osterwalder, A., et al.: Value Proposition Design: How to Create Products and Services Customers Want. Wiley, Hoboken (2014)
6. Ries, E.: The Lean Startup: How Today’s Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. Crown Business, New York (2011)
7. Le, N.-T., Warschat, J.: A new approach to biomimetics and problem solving. In: 21st International Conference on Production Research, ICPR 2011 (2011). <http://publica.fraunhofer.de/documents/N-180679.html>
8. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Meth. Instrum. Comput. **28**(2), 203–208 (1996). doi:[10.3758/BF03204766](https://doi.org/10.3758/BF03204766)

# Bidirectional Deep Learning of Context Representation for Joint Word Segmentation and POS Tagging

Prachya Boonkwan<sup>(✉)</sup> and Thepchai Supnithi

Language and Semantic Technology Lab,  
National Electronics and Computer Technology Center, 112 Phahonyothin Road,  
Khlong Nueng, Khlong Luang 12120, Pathumthani, Thailand  
{prachya.boonkwan, thepchai.supnithi}@nectec.or.th

**Abstract.** Word segmentation and POS tagging are crucial steps for natural language processing. Though deep learning facilitates learning a joint model without feature engineering, it still suffers from unreliable word embedding when words are rare or unknown. We introduce two-level backoff models to which morphological information and character-level contexts are integrated. Experimental results on Thai and Chinese show that our backoff models improve the accuracy of both tasks and excels in OOV recovery.

**Keywords:** Word segmentation · Part-of-speech tagging · Joint tasks · Deep learning · Structured perceptron

## 1 Introduction

Word segmentation and POS tagging are indispensable tasks in natural language processing. In the past two decades, both tasks are taken into account as separate steps, and they have to be learned separately. This practice discards crucial information in the POS level that helps constrain word segmentation, letting it succumb to error propagation from the previous step to the next. Recent studies [13, 22, 24, 25] show that joint models of both tasks improve the overall accuracy but they involve delicate feature engineering, in which a large number of complex features are cherry-picked.

Deep learning [1, 7] minimizes the need of feature engineering by automatically learning features with multiple layer perceptrons and recurrent neural networks. A joint model can now be learned as a pipelined process while retaining all crucial information that helps reduce its search space. Deep learning has been used for learning a wide array of joint task typologies, e.g. word segmentation (WS)+POS tagging [26], WS+POS tagging+lexical normalization [10, 19], WS+NER [17], WS+POS tagging+phrase chunking [14], and WS+POS tagging+dependency parsing [20]. These models are built on top of word embedding (vector representation) learned from a large amount of data.

However, word embedding becomes unreliable when the word is rare or unknown. This issue is prevalent in practical use when (1) the language is morphologically rich, (2) proper names and new words constantly emerge from linguistic and cultural contact, and (3) typos and slangs become popular. We will address this issue by incorporating the notion of *context representations* to the word embedding.

In this paper, we propose context-based backoff models to which morphological information and character-level contexts are integrated (Sect. 2). The intuition of our backoff models is simple: although the feature vector for a rare word is unreliable, its prefix, suffix, and surrounding context are informative and can still be used for predicting the word boundaries and POS tags. We efficiently train our model via the structured perceptron algorithm (Sect. 3). We assessed our backoff models (Sect. 4) with two challenging languages: Chinese and Thai, and we found that our models improve the accuracy of both tasks and particularly excels in OOV recovery.

## 2 Dynamic Neural Architecture

We formulate the joint word segmentation and POS tagging tasks as pipeline tagging, in which word boundaries produced by the first task becomes an input for POS tagging, resulting in nonlinear prediction. This requires the use of dynamic neural architecture because the network of the next task depends on the output of the previous one. Our model offers joint training of two strongly dependent tasks which once had to be trained separately in the traditional tagging approaches. By incorporating multiple layers of perceptrons and nonlinear functions (such as sigmoid, tanh, and ReLU), we can automatically extract useful linguistic features that contribute to the prediction.

We propose the dynamic network architecture in Fig. 1. The first layer extracts character-level  $n$ -gram features from the text. The next layer incorporate the  $n$ -gram features with their surrounding contexts using bidirectional recurrent neural networks (RNNs). The output of this step is the prediction of word boundaries. Next, we assign a feature vector (*embedding*) for each predicted word [1, 15]. We also propose two-level backoff models for word embedding to cope with rare words and the OOV issue. The final output of our joint models is a tag sequence inferred from the graph.

### 2.1 Character-Level $n$ -Gram Embedding

The first step is to extract a sequence of  $n$ -gram tuples out of a given character sequence. For each  $n$ -gram, we assign an index for table lookup and a feature vector. We believe that  $n$ -gram embedding would provide far more useful information than character embedding as they also represent character-level contexts. The  $n$ -gram lookup table is a matrix  $\mathbf{M}_{n\text{gram}} \in \mathbb{R}^{d_{n\text{gram}} \times D_{n\text{gram}}}$ , where there are  $D_{n\text{gram}}$  tuples of  $n$ -grams and  $d_{n\text{gram}}$  dimensions for each  $n$ -gram embedding.



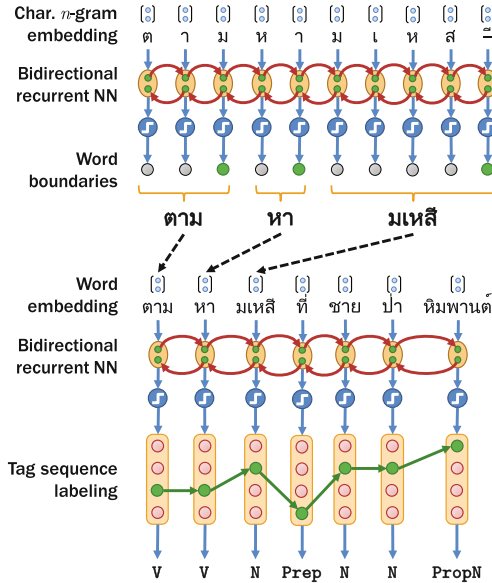


Fig. 1. The dynamic deep network for joint word segmentation and POS tagging tasks

Given an input sequence of characters  $c_{1..n}$ , we first extract an  $n$ -gram tuple  $g_i$  from each character  $c_i$ , where  $g_i = c_{\max(i-n+1,1)..i}$ . We retrieve the  $n$ -gram embedding of each  $g_i$  by the table lookup layer  $G(g_i) = \mathbf{M}_{\text{ngram}} \mathbf{e}_i$ , where  $\mathbf{e}_i$  is a binary vector whose  $i$ -th element is 1 and others are 0.

## 2.2 Word Boundary Inference

The next step is to infer word boundaries from the contexts. To avoid feature co-adaptation that causes the over-fitting issue [23], we randomly drop out the output of some neural units at the drop-out rate  $r_{\text{dropout}}$  while training:

$$\mathbf{g}_i = \text{dropout}(G(g_i), r_{\text{dropout}}) \tag{1}$$

We then compute each LHS context embedding  $\overrightarrow{\mathbf{h}}_{g,i}$  (and RHS context embedding  $\overleftarrow{\mathbf{h}}_{g,i}$ ) by gradually feeding  $n$ -gram embeddings from left to right (and from right to left) into an RNN and retrieving its internal state. Since the RNN computes the output from an input and the previous internal state, it remembers the previous inputs. Therefore, by probing the internal state of the RNN at each step, we are compressing the previously recognized context into a representation vector, which can be used for further prediction.

For each character index  $i$ , we predict the word boundary  $b_i$  by concatenating the contexts from both sides and computing the distribution of segmentation  $\mathbf{b}_i$  by:

$$\begin{aligned} \mathbf{ctx}_{g,i} &= \text{dropout}(\overrightarrow{\mathbf{h}}_{g,i} \oplus \overleftarrow{\mathbf{h}}_{g,i}, r_{\text{dropout}}) \\ \mathbf{b}_i &= \text{softmax}(\mathbf{W}_g f(\mathbf{ctx}_{g,i})) \end{aligned} \quad (2)$$

where  $\oplus$  is the vector concatenation operator,  $f$  is a nonlinear function,  $d_{\text{context}}$  is the dimensions of the context embeddings  $\overrightarrow{\mathbf{h}}_{g,i}$  and  $\overleftarrow{\mathbf{h}}_{g,i}$ , and  $\mathbf{W}_g \in \mathbb{R}^{2 \times 2d_{\text{context}}}$  is a weight matrix projecting to two actions  $\{\text{segment}, \text{append}\}$ . The vector  $\mathbf{b}_i$  for each index  $i$  will be used later in training and inference.

### 2.3 Word Embedding and Two-Level Backoff Models

For POS tagging, we will first assign to each input word a feature vector retrieved by table lookup. Suppose the word lookup table is a matrix  $\mathbf{M}_{\text{word}} \in \mathbb{R}^{d_{\text{word}} \times D_{\text{word}}}$ , where there are  $D_{\text{word}}$  known words, and each feature vector has  $d_{\text{word}}$  dimensions. Given an input sentence  $w_{1..m}$ , we retrieve the feature vector for  $w_i$  by  $\mathbf{w}_i = \mathbf{M}_{\text{word}} \mathbf{e}_i$ , where  $\mathbf{e}_i$  is a binary vector whose  $i$ -th element is 1 and others are 0. The matrix  $\mathbf{M}_{\text{word}}$  can be initialized randomly and automatically trained by any variation of backpropagation, or can be replaced by precomputed word embeddings such as GloVe [18].

We propose the context-based backoff models for word embedding to which morphological information and character-level contexts are integrated. Suppose we encounter an unseen word in Fig. 2(a), we decompose it into characters and  $n$ -gram tuples. In Fig. 2(b), we then learn its prefix and suffix by gradually feeding each character into a bidirectional RNN from both directions to obtain the internal states  $\overrightarrow{\mathbf{h}}_{c,i}$  and  $\overleftarrow{\mathbf{h}}_{c,i}$  at the other ends. To incorporate contextual information from surrounding words, we also learn the prefix and suffix from the  $n$ -gram tuples with another bidirectional RNN and obtain the internal states  $\overrightarrow{\mathbf{h}}_{q,i}$  and  $\overleftarrow{\mathbf{h}}_{q,i}$ . We assume that all of these internal states have the same dimensions  $d_{\text{hidden}}$ .

Finally, we define the feature vector for  $w_i$  interpolated with our backoff models. If word  $w_i$  is rare, i.e.  $\#(w_i) < \theta$  where  $\theta$  is the frequency threshold for backoff, the word embedding for  $w_i$  is

$$\hat{\mathbf{w}}_i = \mathbf{w}_i \oplus \mathbf{W}_{b1}(\overrightarrow{\mathbf{h}}_{c,i} \oplus \overleftarrow{\mathbf{h}}_{c,i}) \oplus \mathbf{W}_{b2}(\overrightarrow{\mathbf{h}}_{q,i} \oplus \overleftarrow{\mathbf{h}}_{q,i}) \quad (3)$$

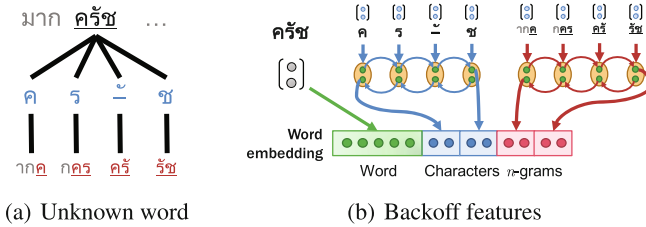


Fig. 2. Context-based backoff models for word embedding

where  $\mathbf{w}_i$  is the word embedding, and  $\mathbf{W}_{b1}, \mathbf{W}_{b2} \in \mathbb{R}^{d_{\text{backoff}} \times 2d_{\text{hidden}}}$  are weight matrices projecting to a backoff vector of  $d_{\text{backoff}}$  dimensions. Otherwise ( $\#(w_i) \geq \theta$ ), the word embedding for  $w_i$  is simply zero-padded on RHS, i.e.  $\hat{\mathbf{w}}_i = \mathbf{w}_i \oplus \mathbf{0}$ . The resultant word embedding  $\hat{\mathbf{w}}_i$  always has  $d_{\text{word}} + 2d_{\text{backoff}}$  dimensions.

## 2.4 Tag Inference

Once we achieve the word embedding with backoffs in Eq. (3), we can now infer a tag sequence from an input sequence of characters. Similar to word segmentation, there are strong dependencies between each word-tag pair and between each consecutive tag pair. This suggests that local and global prediction works side by side. By letting there are  $N_{\text{tags}}$  known tags in the tagset  $T$ , we separate the prediction task into two levels:

**Local Prediction:** For each word-tag pair, we make prediction from the word embedding, its morphological information, and its context by using a multilayer perceptron and a nonlinear function. The prediction score is given below:

$$\begin{aligned} \mathbf{h}_{t,i} &= f(\mathbf{W}_h \times \text{dropout}(\hat{\mathbf{w}}_i, r_{\text{dropout}})) \\ \mathbf{p}_i &= \text{softmax}(\mathbf{W}_o \mathbf{h}_{t,i}) \end{aligned} \quad (4)$$

where  $\mathbf{h}_{t,i}$  is a hidden layer with  $d_{\text{hidden}}$  dimensions,  $f$  is a nonlinear function,  $\mathbf{p}_i \in \mathbb{R}^{N_{\text{tags}}}$  is the output vector whose each element  $p_{i,t_i}$  is the probability of  $w_i$  being tagged as  $t_i \in T$ , and  $\mathbf{W}_h \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{wordrep}}}$  and  $\mathbf{W}_o \in \mathbb{R}^{N_{\text{tags}} \times d_{\text{hidden}}}$  are weight matrices that project to hidden units and tag prediction, respectively. We again drop out some input units to avoid feature co-adaptation.

**Global Prediction:** We take into account the tag transition as a function that takes two consecutive tags  $t_{i-1}$  and  $t_i$ . For the sake of simplicity, we let this score be a matrix  $\mathbf{A} \in \mathbb{R}^{N_{\text{tags}} \times N_{\text{tags}}}$ , whose element  $\alpha_{ji}$  is a score for transitioning from tag  $t_j$  to tag  $t_i$ .

Let us define two quantities: *emission score*  $s_e(i, k) = p_{i,k}$  and *transition score*  $s_t(j, i) = \alpha_{ji}$ . We finally define the score for a possible tag sequence  $t_{1..m}$  given an input sentence  $c_{1..n}$  as follows:

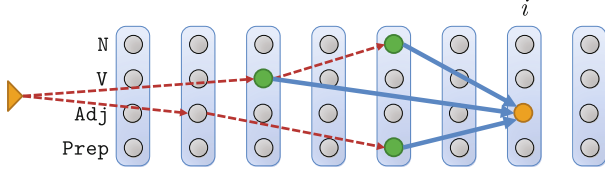
$$s(t_{1..m} | c_{1..n}) = \sum_i [s_e(i, t_i) + s_t(t_{i-1}, t_i)] \quad (5)$$

where  $p_{i,t_i}$  is a prediction score of the current tag  $t_i$  given word embedding  $\hat{\mathbf{w}}_i$ , and  $\alpha_{t_{i-1}, t_i}$  is the transition score from the previous tag  $t_{i-1}$  to  $t_i$ . We can find the best tag sequence  $t_{1..m}^*$  by:  $t_{1..m}^* = \arg \max_{t_{1..m}} s(t_{1..m} | c_{1..n})$ .

## 3 Structured Perceptron Algorithm

### 3.1 Parameter Estimation

In training, we estimate all embedding and weight matrices  $\Theta = (\mathbf{M}_{\text{ngram}}, \mathbf{M}_{\text{word}}, \mathbf{W}_g, \mathbf{W}_{b1}, \mathbf{W}_{b2}, \mathbf{W}_h, \mathbf{W}_o)$  with respect to some objective function over the training dataset. Since our joint tasks are pipeline prediction, we choose to estimate



**Fig. 3.** The search space of our structured perceptron training. Each column corresponds to a character index, while each of its nodes corresponds to a POS tag’s index. Above, there are three choices of forming a word as an adjective (**Adj**) at index  $i$  given the previously explored paths (dashed arrows).

them with the structured perceptron method [6]. In essence, we compute the best path for each input sequence with current parameters, and dynamically create a network along the pipeline. Once we reach the end of the sequence, we update the parameters with backpropagation on the resultant network. By doing so, we are closing the gap between the hypothesis and the training data.

For each entry of the training data, we define the perceptron loss function for an input sequence of characters  $\mathbf{x} = c_{1\dots n}$ , an observed sequence of word boundaries  $\mathbf{B} = \mathbf{b}_{1\dots n}$ , and an observed sequence of POS tags  $\mathbf{t} = t_{1\dots m}$ :

$$\mathcal{L}(\mathbf{x}, \mathbf{B}, \mathbf{t}) = \text{BLL}(\mathbf{B}, \hat{\mathbf{B}}) + \max(s(\hat{\mathbf{t}}|\mathbf{x}) - s(\mathbf{t}|\mathbf{x}), 0) \quad (6)$$

where the current parameters predict a hypothetical word boundary sequence  $\hat{\mathbf{B}}$  and a hypothetical tag sequence  $\hat{\mathbf{t}}$ . The binary log loss of word segmentation between the training  $\mathbf{B}$  and the hypothetical  $\hat{\mathbf{B}}$  can be computed by

$$\text{BLL}(\mathbf{B}, \hat{\mathbf{B}}) = - \sum_{j=1}^n \sum_i \left[ \hat{b}_{ji} \log b_{ji} - (1 - \hat{b}_{ji}) \log (1 - b_{ji}) \right] \quad (7)$$

Finding the globally best hypothesis  $\hat{\mathbf{B}}$  and  $\hat{\mathbf{t}}$  can be done via the Viterbi algorithm. Our search space is illustrated in Fig. 3. Each word-tag pair  $(c_{j\dots i}, t_k)$  is formed by drawing an arrow from the start index  $j - 1$  to the end index  $i$  at tag  $k$ . Since we want to find the maximum difference between the hypothetical tag sequence and the observed one, we assume that this can be approximated by a greedy algorithm:

$$s(\hat{\mathbf{t}}|\mathbf{x}) \approx \sum_{(i,k) \in \hat{\pi}} \sigma(i, k) \quad (8)$$

where  $\hat{\pi}$  is the greedy path of the tag sequence  $\hat{\mathbf{t}}$ . For index  $i \geq 1$ , the path score  $\sigma(i, k)$  for index  $i$  and tag  $k$  is the maximum score of the incoming paths:

$$\sigma(i, k) = s_e(i, k) + \max_{j, k'} [\sigma(j, k') + s_t(k', k)] \quad (9)$$

where  $0 < j < i$ ,  $s_e(i, k)$  is the emission score computed by Eq. (4), and  $s_t(k', k)$  is the transition score from tag  $k'$  to tag  $k$ . Otherwise, we define the initial path score as

$$\sigma(0, k) = \begin{cases} 0 & \text{if } t_k \text{ is the start symbol} \\ -\infty & \text{otherwise} \end{cases} \quad (10)$$

Finally, note that  $L(\mathbf{x}, \mathbf{B}, \mathbf{t})$  is not differentiable due to the hinge loss. We minimize the perceptron loss of the training set  $\mathcal{D}$  with a generalization of gradient descent algorithms called subgradient method [21].

$$\frac{\partial}{\partial \Theta} \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}(\mathbf{x}_i, \mathbf{B}_i, \mathbf{t}_i) = \frac{\partial}{\partial \Theta} \sum_{i=1}^{|\mathcal{D}|} \left[ \text{BLL}(\mathbf{B}_i, \hat{\mathbf{B}}_i) + s(\hat{\mathbf{t}}_i | \mathbf{x}_i) - s(\mathbf{t}_i | \mathbf{x}_i) \right] \quad (11)$$

### 3.2 Joint Inference of Word Segmentation and POS Tagging

Given an input sequence of characters  $\mathbf{x}$ , we infer word boundaries and a tag sequence from the estimated model parameters. We slightly modify the path score to normalize the lengths of words by simply ignoring mid-word character appending. For index  $i \geq 1$ , the path score for decoding becomes:

$$\sigma_{\text{decode}}(i, k) = s_e(i, k) + \max_{j, k'} \left[ \sigma(j, k') + s_t(k', k) + \hat{b}_{j, \text{segment}} + \hat{b}_{i, \text{segment}} \right] \quad (12)$$

where  $\hat{b}_{j, \text{segment}}$  and  $\hat{b}_{i, \text{segment}}$  are the probabilities of segmenting action at indices  $j$  and  $i$ , respectively, computed by Eq. (2). Otherwise, the initial path score is  $\sigma_{\text{decode}}(0, k) = \sigma(0, k)$  as described in Eq. (10).

**Postprocessing:** We find that some rare or unknown words are written in a foreign script. This makes character embedding and  $n$ -gram embedding also unreliable due to data sparseness. After decoding, we find that each foreign-language character is segmented as a singleton. We can easily overcome this issue by combining consecutive singletons written in a foreign script, including mid-word spaces, as one unit and assign a POS tag. By our intuition, we assign all discovered foreign words to a noun.

## 4 Experiments

### 4.1 Settings

**Datasets:** We choose two datasets: Thai Orchid-2 and Penn Chinese Treebank (PCTB) [4], to represent Thai and Chinese. For Thai, Orchid-2 consists of texts from Thai newspaper and magazines, and all sentences are manually sentence-segmented, word-segmented and POS-tagged. We randomly select 10% of all sentences as the testing set, leaving the remaining 90% as the training set. For Chinese, PCTB consists of texts from Chinese newspaper and broadcast news, which are manually word-segmented, POS-tagged, and annotated with syntactic structures. We remove the syntactic structures from all trees to obtain their preterminal nodes. Both datasets have different guidelines: Thai Orchid-2 prefers

**Table 1.** Statistics of the datasets

	Numbers	Lengths				Numbers	Lengths		
		Min	Max	Avg			Min	Max	Avg
<i>(a) Thai Orchid-2</i>					<i>(b) Penn Chinese Treebank (PCTB)</i>				
Training sents	8,535	1	95	13.3	Training sents	56,957	1	50	6.2
Testing sents	949	1	57	13.0	Testing sents	867	1	29	6.0
Known words	11,215	1	71	27.4	Known words	37,768	1	22	2.4
Unknown words	613	1	33	15.7	Unknown words	327	1	10	2.7
Tagset	50	—			Tagset	203	—		

small morphemes while Chinese PCTB prefers large compound words. Both of them contain a foreign script, i.e. Roman Alphabet and Arabic numerals. The statistics for both datasets are shown in Table 1.

**Baselines:** We compare our model with the following baselines. The baselines for Chinese are the state-of-the-art joint models [13, 14, 19, 24–26]. For Thai, we compare our model with existing word segmentation techniques [11, 12] and POS tagging methods [2, 16] trained on BEST Corpus [3]. We also compare our model with CRF-based joint pointwise models of word segmentation and POS tagging. Avoiding extreme feature engineering, we choose to assess the performance of this baseline with the rather simple features as follows: surrounding character trigrams and POS tag bigrams. We train it with the L-BFGS Algorithm for 20 epochs.

**Our model:** For the sake of experimental simplicity, we reduce the number of variables. Excluding the word embedding whose dimension is  $d_{\text{word}}$ , we equate the dimensions of all embedding, hidden layers and internal states of RNNs in our model to the dimensions of backoff ( $d_{\text{backoff}}$ ). We use GRUs [5] as our RNNs throughout the model as they perform almost as well as LSTMs [9] with less computational complexity. We set the frequency threshold for backoff  $\theta = 5$ . All dropout units share the same dropout rates  $r_{\text{dropout}} = 0.5$ . We initialize the model parameters with Glorot’s [8] method. We train our model with the vanilla Stochastic Gradient Descent algorithm with learning rate  $\eta = 0.1$ . In training, we separate the training set into mini-batches of various sizes to boost the speed of convergence. We also disable and enable the postprocessing in decoding to observe its before-and-after effects.

**Performance Measurement:** We measure and compare the performance of our joint model with the baselines via the balanced F1 score:  $F_1 = \frac{2PR}{P+R}$ , where  $P$  is the precision and  $R$  is the recall. We also employ the OOV recall for comparing the performance of unknown word recovery. The performance of each task will be denoted by the following abbreviations: WS for word segmentation F1, Tag for POS tagging F1, and OOV for unknown word recall. If the postprocessing is enabled, we also add the suffix ‘+Post’ to these abbreviations, such as WS + Post.

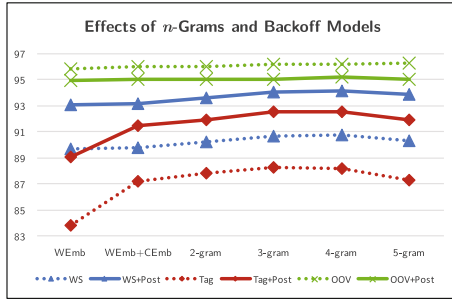
**Evaluation Scheme:** We will compare the best performances of our model and the baselines. We are fully aware that some of our baselines were evaluated on smaller datasets such as Chinese Treebanks 4, 5, and 7 (CTB-4, -5, and -7). However, since we do not possess any of them, we will instead present our model’s performance on PCTB for an analogy. Next, we will evaluate our model’s behaviors in depth on Thai Orchid-2, because it is meticulous to observe the prediction’s accuracy on longer word lengths.

## 4.2 Multilingual Results

We present the performance of our models and the baselines in Table 2. To the best of our knowledge, we are the first who report the results of the joint tasks for word segmentation and POS tagging for Thai. Since these models are trained on different datasets, we cannot directly compare them head to head. However, our model still yields relatively accurate results (above 90%) on word segmentation F1 (WS), POS tagging F1 (Tag), and unknown word recall (OOV) on our Chinese and Thai datasets. When evaluated on our datasets, our model significantly outperforms our CRF-based pointwise baseline. This is because structured perceptron helps reduce the search space with its Viterbi-styled training.

**Table 2.** Best multilingual results on the joint tasks of word segmentation and POS tagging. WS stands for word segmentation F1, Tag for POS tagging F1, and OOV for unknown-word recall. The settings of our best model for Chinese are:  $n = 2$ ,  $d_{\text{word}} = 300$ ,  $d_{\text{backoff}} = 500$ ,  $\theta = 5$ ,  $r_{\text{dropout}} = 0.5$ , and the mini-batch size = 5. The settings for Thai are:  $n = 3$ ,  $d_{\text{word}} = 100$ ,  $d_{\text{backoff}} = 300$ ,  $\theta = 5$ ,  $r_{\text{dropout}} = 0.5$ , and the mini-batch size = 5. Note that Murata et al. [16] reported only the precision on their small handcrafted dataset.

Models	Chinese				Thai			
	Datasets	WS	Tag	OOV	Datasets	WS	Tag	OOV
Kruengkrai et al. [13]	CTB-5	97.87	93.67	—				
Zhang and Clark [25]	CTB-5	95.84	91.37	—				
Zeng et al. [24]	CTB-7	96.85	92.89	68.09				
Zheng et al. [26]	CTB-4	95.23	72.38	91.82				
Qian et al. [19]	CTB-7	91.62	84.01	—				
Lyu et al. [14]	CTB-4	90.67	82.45	—				
Kruengkrai et al. [12]					BEST	97.84	—	—
Kongyoung et al. [11]					BEST	97.64	—	—
Murata et al. [16]					?	93.90*		
Boonkwan et al. [2]					BEST	—	97.62	—
Our pointwise CRF	PCTB	78.13	56.59	63.99	Orchid-2	74.86	67.58	46.58
<b>Ours</b>	PCTB	94.02	91.97	<b>99.53</b>	Orchid-2	90.81	88.66	<b>96.31</b>
<b>Ours + Post</b>	PCTB	<b>94.13</b>	<b>92.11</b>	99.23	Orchid-2	<b>94.18</b>	<b>92.78</b>	95.23



**Fig. 4.** Effects of  $n$ -grams and backoff models ( $d_{\text{word}} = d_{\text{backoff}} = 100$ ,  $\theta = 5$ , batch size = 5)

### 4.3 Effects of $n$ -Grams and Backoff Models

Henceforth, we will investigate the behaviors of our model in details on Thai Orchid-2. We examine the effects of  $n$ -grams and the efficiency of our backoff models for the word embedding. We fix the dimensions of word embedding ( $d_{\text{word}}$ ) and context embedding ( $d_{\text{backoff}}$ ) to 100, the mini-batch size to 5, and the backoff threshold to 5. In Fig. 4, we vary the number of  $n$ -grams from 0-grams (using only word embedding; WEmb) and unigrams (using word and character embeddings; WEmb + CEmb) up to 5-grams, and observe the F1 scores and the OOV recalls.

The results reveal that the use of  $n$ -grams slightly improves the word segmentation F1 (+1.5%) and significantly leverages the POS tagging F1 (+4%), while not affecting the OOV recall. We understand that this increase is due to the inclusion of context characters in the  $n$ -gram backoff level. But as we increase the grams, both F1 scores start to drop. We imply that this is due to data sparseness as larger  $n$ -grams are more sensitive to noise. When switching on the postprocessing, both F1 scores increase by 2% each but the OOV recall drops by 1%. We suspect that our intuition of combining consecutive character singletons may be an over-generalization, and it should be replaced by more accurate classifiers.

### 4.4 Effects of Word Embedding and Context Embedding

Next, we investigate the effects of the layer dimensions of word embedding ( $d_{\text{word}}$ ) and context embedding ( $d_{\text{backoff}}$ ). We use trigrams and set the mini-batch size to 5 and the backoff threshold to 5. We then vary the layer dimensions of the word embedding and the context embedding.

To observe the effects of the word embedding, we vary the layer dimensions of the word embedding between 25, 50, 100, 300, 500, and 1,000. In this experiment, we shrink the context embedding to 10 dimensions to minimize its effects on the prediction. The results in Fig. 5(a) show that the F1 scores increase and saturate very quickly at 50 dimensions. Turning on the postprocessing significantly boosts



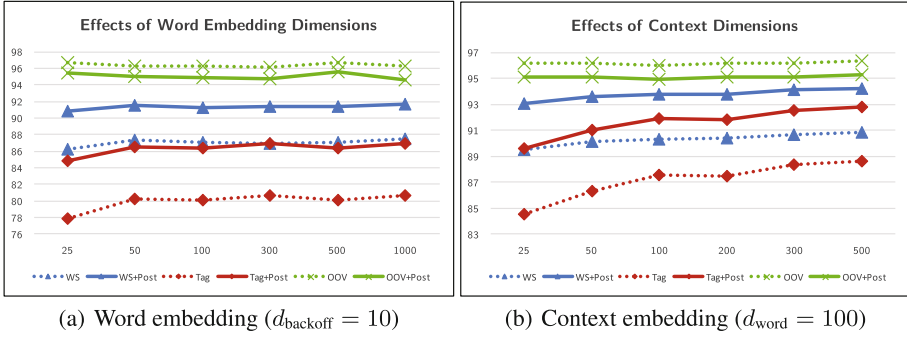


Fig. 5. Effects of layer dimensions ( $\theta = 5$ , batch size = 5)

the F1 scores for both word segmentation (+4%) and POS tagging (+6.25%). However, we also found that setting the context embedding to higher dimensions (e.g. 100) diminishes this performance increase. It suggests that moderate dimensions of word embedding ( $d_{word}$ ) would suffice to represent known words.

We further vary the layer dimensions of the context embedding between 25, 50, 100, 300, and 500. In this experiment, we set the dimensions for word embedding to 100. Significantly, in Fig. 5(b), we found that higher dimensions gradually increase the F1 scores for word segmentation (+4%) and POS tagging (+4.25%). Switching on the postprocessing boosts both F1 scores by 4 more percents. We imply that we can expand the context embedding to cope with unknown words accurately.

#### 4.5 Effects of Mini-Batch Sizes

Finally, we observe the effects of the mini-batch sizes. We use trigrams, set the backoff threshold to 5, and set the dimensions of word embedding and context embedding to 100 and 300, respectively. We vary the size of mini-batches between 1, 2, 5, 10, 20, 50, and 100. The results are shown in Fig. 6.

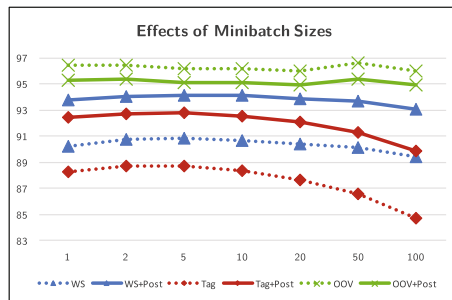


Fig. 6. Effects of mini-batch sizes ( $d_{word} = 100$ ,  $d_{backoff} = 300$ ,  $\theta = 5$ )

We found that by increasing the mini-batch size, both F1 scores slightly improve and start to plunge—word segmentation by  $-2\%$  and POS tagging by  $-4\%$ . We imply that the convergence of our model depends on the number of parameter updates rather than the number of epochs. For large mini-batch sizes, parameters updates are less frequent than those of smaller mini-batch ones. Both F1 scores slightly drop at small mini-batch sizes (1 and 2) because parameter updates take place slightly too often, causing the model to overfit. Adding a parameter-regularizing criterion such as  $L_1$  and  $L_2$  would solve this issue. Switching on the postprocessing increases both F1 scores by  $+2\%$ .

## 5 Conclusion

We have presented a dynamic neural network architecture for the joint tasks of word segmentation and POS tagging, and introduced two-level backoff models based on context representations to cope with rare words and the OOV issue. Our model is learned via the structured perceptron algorithm that retains crucial information in both tasks and reduces the search space. The results reveal that our backoff models significantly improve the overall accuracy of word segmentation, POS tagging, and OOV recovery. The more dimensions the context embedding has, the more accuracy we will achieve.

Our future work is as follows. First, we will investigate why the dimensions of word embedding has much less effects on the overall accuracy than the context embedding. Second, we will replace the postprocessing step with more accuracy classifiers. Third and finally, we will extend our network architecture to the task of sentence segmentation, because this step is indispensable for Thai NLP.

## References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. *JMLR* **3**, 1137–1155 (2003)
2. Boonkwan, P., Supnithi, T., Pailai, J., Kongkachandra, R.: Gradient-descent error correction of POS tagging. In: *Proceedings of SNLP* (2013)
3. Boriboon, M., Kriengkiet, K., Chootrakool, P., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T., Kosawat, K.: BEST corpus development and analysis. In: *Proceedings of the 2009 International Conference on Asian Language Processing*, pp. 322–327 (2009)
4. Chen, K.L., Hsieh, Y.M.: Chinese treebanks and grammar extraction. In: *Proceedings of IJCNLP*, pp. 560–565 (2004)
5. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: *NIPS 2014: Deep Learning and Representation Learning Workshop* (2014)
6. Collins, M.: Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In: *Proceedings of EMNLP* (2002)
7. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.P.: Natural language processing (almost) from scratch. *JMLR* **12**, 2493–2537 (2011)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of AISTats*, vol. 9, pp. 249–256 (2010)

9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
10. Kaji, N., Kitsuregawa, M.: Accurate word segmentation and POS tagging for japanese microblogs: corpus annotation and joint modeling with lexical normalization. In: *Proceedings of EMNLP*, pp. 99–109 (2014)
11. Kongyoung, S., Rugchatjaroen, A., Kosawat, K.: TLex+: a hybrid method using conditional random fields and dictionaries for Thai word segmentation. In: *Proceedings of KICSS* (2015)
12. Kruengkrai, C., Uchimoto, K., Kazama, J., Torisawa, K., Isahara, H., Jaruskulchai, C.: A word and character-cluster hybrid model for Thai word segmentation. In: *Proceedings of InterBEST 2009: Thai Word Segmentation Workshop*, pp. 24–29 (2009)
13. Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., Isahara, H.: An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In: *Proceedings of the Joint Conference of the 47th ACL and the 4th IJCNLP of the AFNLP*, vol. 1, pp. 513–521 (2009)
14. Lyu, C., Zhang, Y., Ji, D.: Joint word segmentation, POS-tagging, and syntactic chunking. In: *Proceedings of AAAI*, pp. 3007–3014 (2016)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Proceedings of NIPS* (2013)
16. Murata, M., Ma, Q., Isahara, H.: Part of speech tagging in Thai language using support vector machine. In: *Proceedings of NLPRS: The 2nd Workshop on Natural Language Processing and Neural Networks* (2001)
17. Peng, N., Dredze, M.: Improving named entity recognition for Chinese social media with word segmentation representation learning. In: *Proceedings of ACL*, pp. 149–155 (2016)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: *Proceedings of EMNLP*, pp. 1532–1543 (2014)
19. Qian, T., Zhang, Y., Zhang, M., Ren, Y., Ji, D.: A transition-based model for joint segmentation, POS-tagging, and normalization. In: *Proceedings of EMNLP*, pp. 1837–1846 (2015)
20. Qian, X., Liu, Y.: Joint Chinese word segmentation, POS tagging, and parsing. In: *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, pp. 501–511 (2012)
21. Ratliff, N., Bagnell, J.A., Zinkevich, M.: (Online) Subgradient methods for structured prediction. In: *Proceedings of AISTats* (2007)
22. Shi, Y., Wang, M.: A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks. In: *Proceedings of the IJCAI*, pp. 1707–1712 (2007)
23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *JMLR* **15**, 1929–1958 (2014)
24. Zeng, X., Wong, D.F., Chao, L.S., Trancoso, I.: Graph-based semi-supervised model for joint Chinese word segmentation and part-of-speech tagging. In: *Proceedings of ACL*, pp. 770–779 (2013)
25. Zhang, Y., Clark, S.: A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In: *Proceedings of EMNLP*, pp. 843–852 (2010)
26. Zheng, X., Chen, H., Xu, T.: Deep learning for Chinese word segmentation and POS tagging. In: *Proceedings of EMNLP*, pp. 647–657 (2013)

# A Model for a Computing Cluster with Two Asynchronous Servers

Hai T. Nguyen<sup>2</sup> and T.V. Do<sup>1,2</sup>(✉)

<sup>1</sup> Division of Knowledge and System Engineering for ICT,  
Faculty of Information Technology, Ton Duc Thang University,  
Ho Chi Minh City, Vietnam

dovantien@tdt.edu.vn

<sup>2</sup> Analysis, Design and Development of ICT Systems (AddICT) Laboratory,  
Department of Networked Systems and Services,  
Budapest University of Technology and Economics,  
Magyar tudósok körútja 2, Budapest 1117, Hungary  
do@hit.bme.hu

**Abstract.** Queues with vacations can be used to model computing systems where mechanisms are applied to save the energy consumption of servers in computing clusters. This paper provides the analysis of a cluster with two asynchronous server with the use of the  $M/M/2$  queue with working vacations.

## 1 Introduction

Queues with vacations can be used to model computing systems where mechanisms are applied to save the energy consumption of servers in computing clusters. In the term of queueing theory, server vacations represents an event that physical servers go into a sleeping state if no processing needs can be found in the system.

Servi and Finn [6] introduced a variation of the vacation queues called the working vacations. In this case when the traffic is low the server can enter a working vacation state in which it is still available but operates with a lower performance. Their study was later extended by Baba [2] who worked on the  $GI/M/1$  with multiple working vacations. Chae et al. [3] compared the  $GI/M/1$  to the  $GI/Geo/1$  both with single working vacations. Zhang and Tian [8,9] studied the  $M/M/c$  queue with synchronous vacations. Altman and Yechiali [1] compared the  $M/M/c$  with single vacation to its multiple variant. Lin and Ke [4] considered the  $M/M/c$  with single working vacations. Recently, Wang et al. [7] studied multi-server systems with working vacations, impatient customers and they considered synchronous state changes. To extend their works, we consider a computing cluster where the operation of two servers is asynchronous.

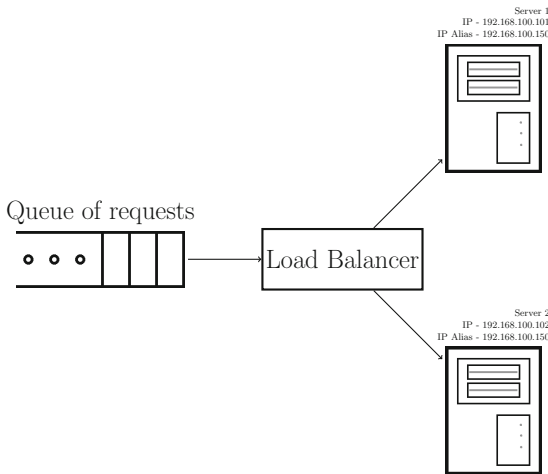
The rest of this paper is organized as follows. In Sect. 2 we present an analytical model and analysis for a two-server configuration with Pacemaker. In Sect. 3 we conduct a numerical analysis and compute the average power and the average waiting time of different scenarios for comparison.

## 2 Asynchronous Working Vacations

When a physical computer suffers a breakdown all the service hosted by it halts until the computer is restarted. Such failures can cause longer outages that are not affordable in modern ICT systems. As a remedy it is recommended to run services on multiple computers connected into a High Availability (HA) cluster.

We consider a computing cluster the two servers change states independently (see Fig. 1). Such a cluster configuration is commonly established in enterprise environment and small companies. To enhance the reliability two servers are configured with a resource manager (e.g., Pacemaker<sup>1</sup>) that allows the configuration and the operation of the cluster. It supports both active-passive and active-active configurations. In an active-passive setup services run on only one computer and upon failure another computer takes over. In case of an active-active scenario services run on all computer nodes and the incoming requests are distributed through load balancing.

In Fig. 1, we can see an active-active configuration with two nodes. Here the two computers are under the same IP alias which is made possible through the IPAddr2 resource agent. The agent runs with a clone on both nodes and every time a request arrives on the IP alias the load balancer sends it to one of the machines. Since the two computers look identical from the outside they also need to have the same resources installed on them. In case one of the nodes fail all requests are directed towards the other one until Pacemaker detects and restores the failed node.



**Fig. 1.** An active-active setup

After the server finishes its job if it cannot find a customer waiting for service it goes into a vacation period regardless of whether the other server is still serving

<sup>1</sup> <http://clusterlabs.org/>.

a request or not. The duration of the vacation, the rate of arrival and the service rate in the normal states and the vacation states are exponentially distributed with  $\vartheta, \lambda, \mu_b$  and  $\mu_v$ , respectively.

Let  $I(t) \in \{0, 1, 2\}$  be the number of servers in vacation and let  $J(t) = \{0, 1, 2, \dots\}$  be the number of customers in the system at time  $0 < t \in \mathbb{R}$ . The system can be described by Markov chain  $\{I(t), J(t)\}$  over the state space  $\mathbb{S} = \{0, 1, 2\} \times \{0, 1, 2, \dots\}$ .

### 2.1 The Steady State Probabilities

We can write the balance equations as

$$\mathbf{Q}_0 \mathbf{v}_{j-1} + \mathbf{Q}_1 \mathbf{v}_j + \mathbf{Q}_2 \mathbf{v}_{j+1} = 0, \quad j \geq 3, \tag{1}$$

where

$$\begin{aligned} \mathbf{Q}_0 &= \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \\ \mathbf{Q}_1 &= \begin{bmatrix} -(\lambda + 2\mu_b) & \vartheta & 0 \\ 0 & -(\vartheta + \lambda + \mu_v + \mu_b) & 2\vartheta \\ 0 & 0 & -(2\vartheta + \lambda + 2\mu_v) \end{bmatrix}, \\ \mathbf{Q}_2 &= \begin{bmatrix} 2\mu_b & 0 & 0 \\ 0 & \mu_b + \mu_v & 0 \\ 0 & 0 & 2\mu_v \end{bmatrix}. \end{aligned}$$

For the case of  $j < 3$  we can write

$$\lambda p(0, 0) = \vartheta p(1, 0), \tag{2}$$

$$(\vartheta + \lambda)p(1, 0) = 2\vartheta p(2, 0) + \mu_b p(0, 1), \tag{3}$$

$$(2\vartheta + \lambda)p(2, 0) = \mu_b p(1, 1) + \mu_v p(2, 1), \tag{4}$$

$$(\mu_b + \lambda)p(0, 1) = \lambda p(0, 0) + \vartheta p(1, 1), \tag{5}$$

$$\begin{aligned} (\vartheta + \mu_b + \lambda)p(1, 1) &= \lambda p(1, 0) + 2\vartheta p(2, 1) \\ &\quad + 2\mu_b p(0, 2) + \mu_v p(1, 2), \end{aligned} \tag{6}$$

$$\begin{aligned} (\mu_v + 2\vartheta + \lambda)p(2, 1) &= \lambda p(2, 0) + \mu_b p(1, 2) \\ &\quad + 2\mu_v p(2, 2), \end{aligned} \tag{7}$$

$$(2\mu_b + \lambda)p(0, 2) = \lambda p(0, 1) + \vartheta p(1, 2) + 2\mu_b p(0, 3). \tag{8}$$

We look for the solutions  $\mathbf{v}_j$  of (1) in the form of

$$\mathbf{v}_j = x_1 \lambda_1^j \boldsymbol{\psi}_1 + x_2 \lambda_2^j \boldsymbol{\psi}_2 + x_3 \lambda_3^j \boldsymbol{\psi}_3. \tag{9}$$

with  $\lambda_i$  being the eigenvalues and  $\boldsymbol{\psi}_i$  being the eigenvectors of the system (see [5]).

**Theorem 1.** *The eigenvalues of the system inside the unit circle are*

$$\begin{aligned} \lambda_1 &= \frac{\lambda}{2\mu_b}, \\ \lambda_2 &= \frac{2\vartheta + \lambda + 2\mu_v - \sqrt{(2\vartheta + \lambda + 2\mu_v)^2 - 8\lambda\mu_v}}{4\mu_v}, \\ \lambda_3 &= \frac{\vartheta + \lambda + \mu_b + \mu_v - \sqrt{(\vartheta + \lambda + \mu_b + \mu_v)^2 - 4\lambda(\mu_b + \mu_v)}}{2(\mu_b + \mu_v)}. \end{aligned}$$

*Proof.* To get the eigenvalues of the system we have to solve the equation  $\det(\mathbf{Q}(x)) = 0$  and should only consider the solutions inside the unit circle. Notice that

$$\mathbf{Q}(x) = \mathbf{Q}_0 + \mathbf{Q}_1x + \mathbf{Q}_2x^2 = \begin{bmatrix} q_{11} & q_{12} & 0 \\ 0 & q_{22} & q_{23} \\ 0 & 0 & q_{33} \end{bmatrix},$$

where

$$\begin{aligned} q_{11} &= 2\mu_b x^2 - (\lambda + 2\mu_b)x + \lambda, \\ q_{12} &= \vartheta x, \\ q_{22} &= (\mu_b + \mu_v)x^2 - (\vartheta + \lambda + \mu_v + \mu_b)x + \lambda, \\ q_{23} &= 2\vartheta x, \\ q_{33} &= 2\mu_v x^2 - (2\vartheta + \lambda + 2\mu_v)x + \lambda, \end{aligned}$$

which means that  $\mathbf{Q}(x)$  is an upper triangular matrix and that the determinant is a product of three quadratic polynomials:

$$\det(\mathbf{Q}(x)) = q_{11}q_{22}q_{33}.$$

The solutions will be the roots of the polynomials. The root of  $q_{11}$  are

$$\begin{aligned} \eta_1 &= 1, \\ \eta_2 &= \frac{\lambda}{2\mu_b}, \end{aligned}$$

the roots of  $q_{22}$  are

$$\begin{aligned} \eta_3 &= \frac{2\vartheta + \lambda + 2\mu_v - \sqrt{(2\vartheta + \lambda + 2\mu_v)^2 - 8\lambda\mu_v}}{4\mu_v}, \\ \eta_4 &= \frac{2\vartheta + \lambda + 2\mu_v + \sqrt{(2\vartheta + \lambda + 2\mu_v)^2 - 8\lambda\mu_v}}{4\mu_v}, \end{aligned}$$

and the roots of  $q_{33}$  are

$$\eta_5 = \frac{\vartheta + \lambda + \mu_b + \mu_v - \sqrt{(\vartheta + \lambda + \mu_b + \mu_v)^2 - 4\lambda(\mu_b + \mu_v)}}{2(\mu_b + \mu_v)},$$

$$\eta_6 = \frac{\vartheta + \lambda + \mu_b + \mu_v + \sqrt{(\vartheta + \lambda + \mu_b + \mu_v)^2 - 4\lambda(\mu_b + \mu_v)}}{2(\mu_b + \mu_v)}.$$

Obviously  $\eta_1$  cannot be inside the unit circle. We can also prove that  $\eta_4 > 1$  and  $\eta_6 > 1$ . Suppose indirectly that  $\eta_4 < 1$ . We get that

$$2\vartheta + \lambda + 2\mu_v + \sqrt{(2\vartheta + \lambda + 2\mu_v)^2 - 8\lambda\mu_v} < 4\mu_v,$$

$$(2\vartheta + \lambda + 2\mu_v)^2 - 8\lambda\mu_v < (2\mu_v - 2\vartheta - \lambda)^2,$$

$$\mu_v\vartheta < 0,$$

which is a contradiction as both  $\mu_v$  and  $\vartheta$  are positive numbers. Using similar steps we can prove that  $\eta_6 > 1, \eta_3 < 1$  and  $\eta_5 < 1$ . For  $\eta_2 < 1$  to hold true we need  $\lambda < 2\mu_b$ .

The eigenvectors corresponding to the eigenvalues  $\lambda_1, \lambda_2$  and  $\lambda_3$  are

$$\psi_1 = [1 \ 0 \ 0]^T,$$

$$\psi_2 = \left[ 1 - \frac{1}{\vartheta\lambda_2}(\lambda - (\lambda + 2\mu_b)\lambda_2 + 2\mu_b\lambda_2^2) \right. \\ \left. \frac{1}{2(\vartheta\lambda_2)^2}(\lambda - (\lambda + 2\mu_b)\lambda_2 + 2\mu_b\lambda_2^2)(\lambda - (\vartheta + \lambda + \mu_b + \mu_v)\lambda_2 + (\mu_b + \mu_v)\lambda_2^2) \right]^T,$$

$$\psi_3 = \left[ 1 - \frac{1}{\vartheta\lambda_3}(\lambda - (\lambda + 2\mu_b)\lambda_3 + 2\mu_b\lambda_3^2) \ 0 \right]^T.$$

## 2.2 Performance Measures

**Average Queue Length.** The average queue length can be calculated using the formula

$$\text{AvgQ} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} j \cdot p(i, j).$$

We can use (9) to obtain

$$\text{AvgQ}_{\text{async}} = \sum_{i=0}^2 p(i, 1) + \sum_{i=0}^2 \sum_{k=1}^3 x_k \frac{(2 - \lambda_k)\lambda_k^2}{(1 - \lambda_k)^2} \psi_{k,i}. \tag{10}$$

**Average Power.** Let  $P_0$  be the server’s power during normal period and  $P_1$  be the power during a vacation. The average power in the synchronous case therefore is

$$\text{AvgP}_{\text{sync}} = P_0 \sum_{j=0}^{\infty} p(0, j) + P_1 \sum_{j=0}^{\infty} p(1, j)$$

$$= \sum_{l=0}^1 P_l \left( p(l, 0) + \sum_{k=1}^2 x_k \frac{\lambda_k}{1 - \lambda_k} \psi_{k,l} \right). \tag{11}$$



The average power is

$$\begin{aligned}
 \text{AvgP}_{\text{async}} &= 2P_0 \sum_{j=0}^{\infty} p(0, j) + (P_0 + P_1) \sum_{j=0}^{\infty} p(1, j) + 2P_1 \sum_{j=0}^{\infty} p(2, j) \\
 &= 2P_0 \left( p(0, 0) + p(0, 1) + \sum_{k=1}^3 x_k \frac{\lambda_k^2}{1 - \lambda_k} \psi_{k,0} \right) \\
 &\quad + (P_0 + P_1) \left( p(1, 0) + p(1, 1) + \sum_{k=1}^3 x_k \frac{\lambda_k^2}{1 - \lambda_k} \psi_{k,1} \right) \\
 &\quad + 2P_1 \left( p(2, 0) + p(2, 1) + \sum_{k=1}^3 x_k \frac{\lambda_k^2}{1 - \lambda_k} \psi_{k,2} \right).
 \end{aligned} \tag{12}$$

### 3 Numeric Results

For our numeric analysis we examined the average queue length and the average power for various parameter values. When calculating the average power we set  $P_0 = 24.5$  W and  $P_1 = 6$  W. The service rate is  $\mu_b = 5$  in the normal state and  $\mu_v = 1.25$  in the working vacation state.

Figures 2, 3, 8 and 9 show the queue length and Figs. 5, 6, 11 and 12 show the average power against  $\vartheta$  for  $\lambda \in \{0.5, 1.5, 3.5, 5, 7, 9\}$  in the synchronous and the

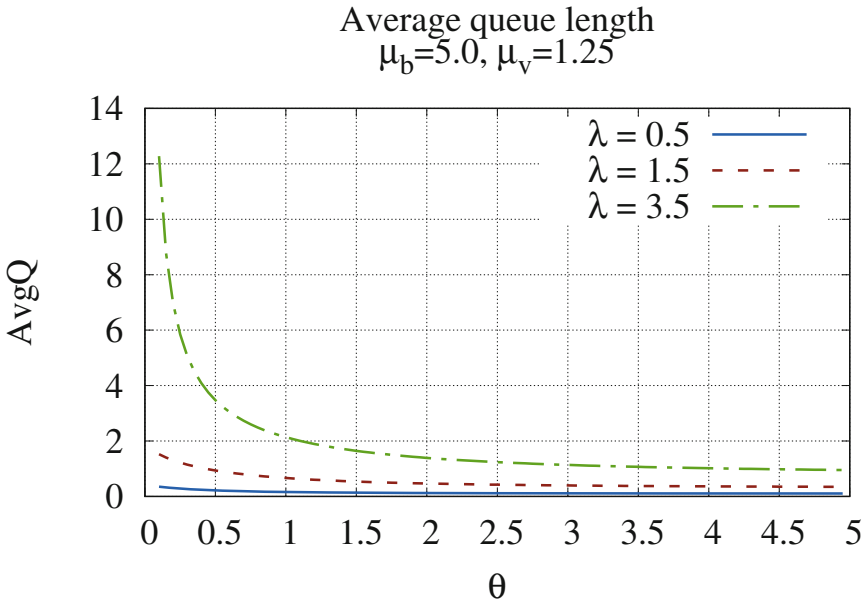


Fig. 2. Average queue length against  $\vartheta$  in the synchronous case for  $\lambda \in \{0.5, 1.5, 3.5\}$ .

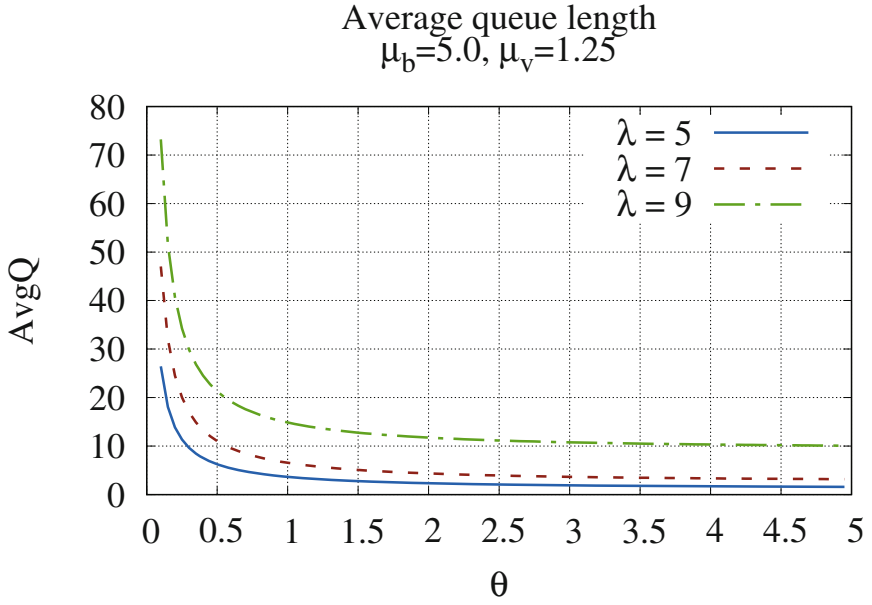


Fig. 3. Average queue length against  $\vartheta$  in the synchronous case for  $\lambda \in \{5, 7, 9\}$ .

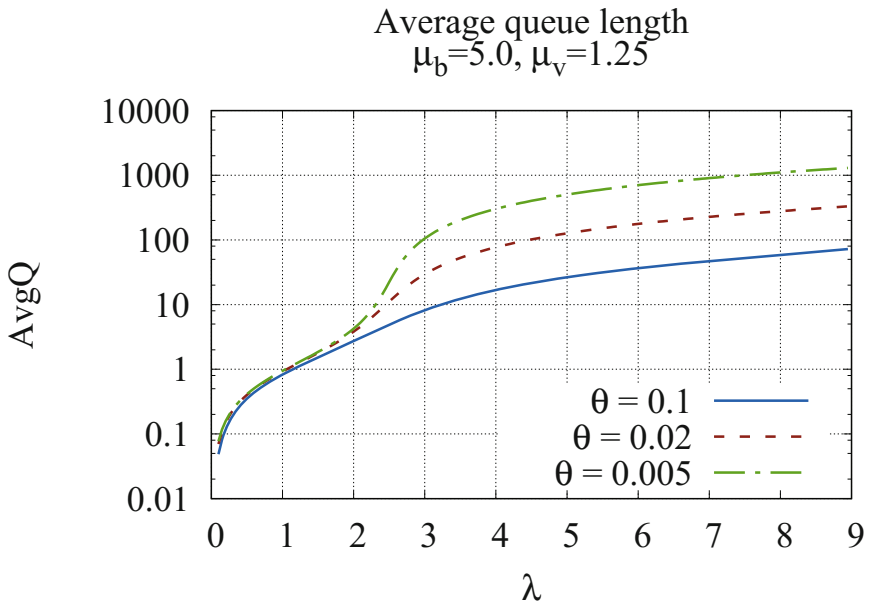


Fig. 4. Average queue length against  $\lambda$  in the synchronous case.

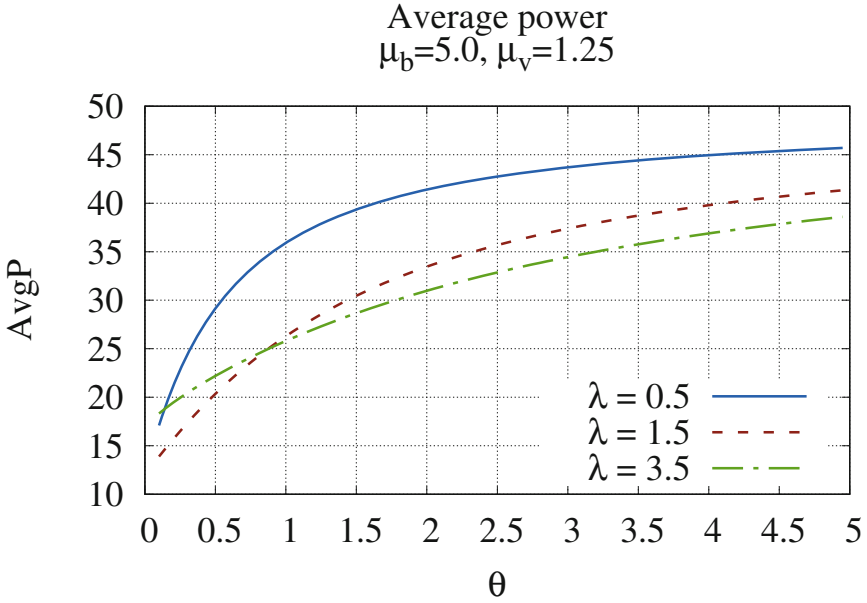


Fig. 5. Average power against  $\vartheta$  in the synchronous case for  $\lambda \in \{0.5, 1.5, 3.5\}$ .

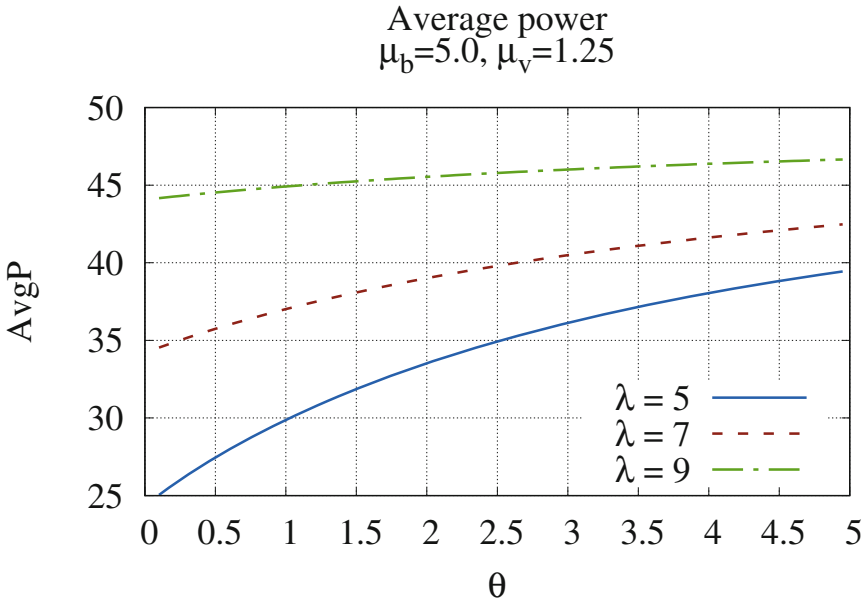


Fig. 6. Average power against  $\vartheta$  in the synchronous case for  $\lambda \in \{5, 7, 9\}$ .

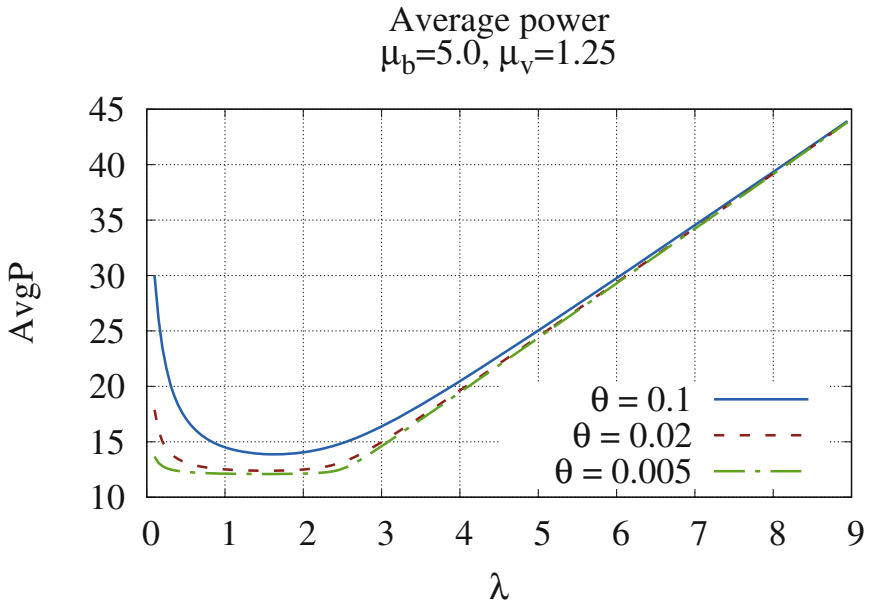


Fig. 7. Average power against  $\lambda$  in the synchronous case.

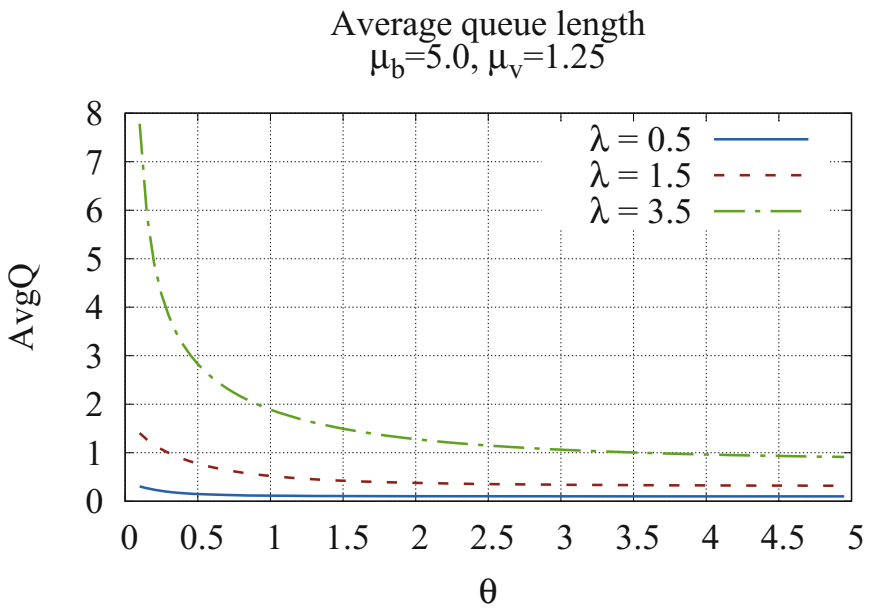


Fig. 8. Average queue length against  $\vartheta$  in the asynchronous case for  $\lambda \in \{0.5, 1.5, 3.5\}$ .

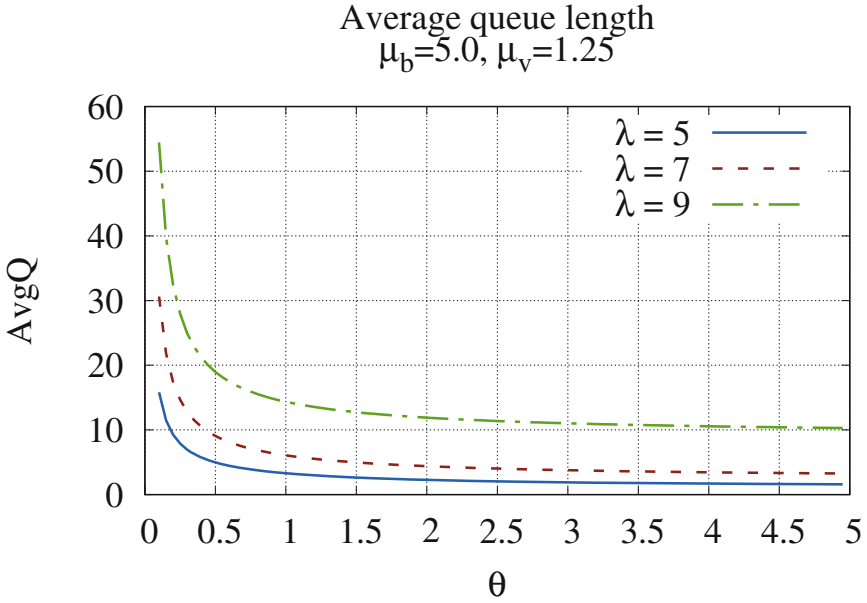


Fig. 9. Average queue length against  $\vartheta$  in the asynchronous case for  $\lambda \in \{5, 7, 9\}$ .

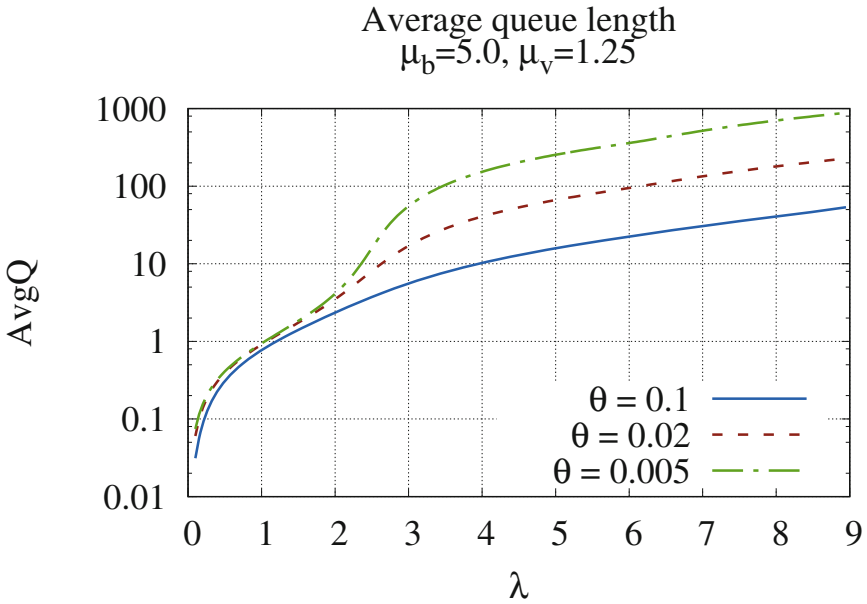


Fig. 10. Average queue length against  $\lambda$  in the asynchronous case.

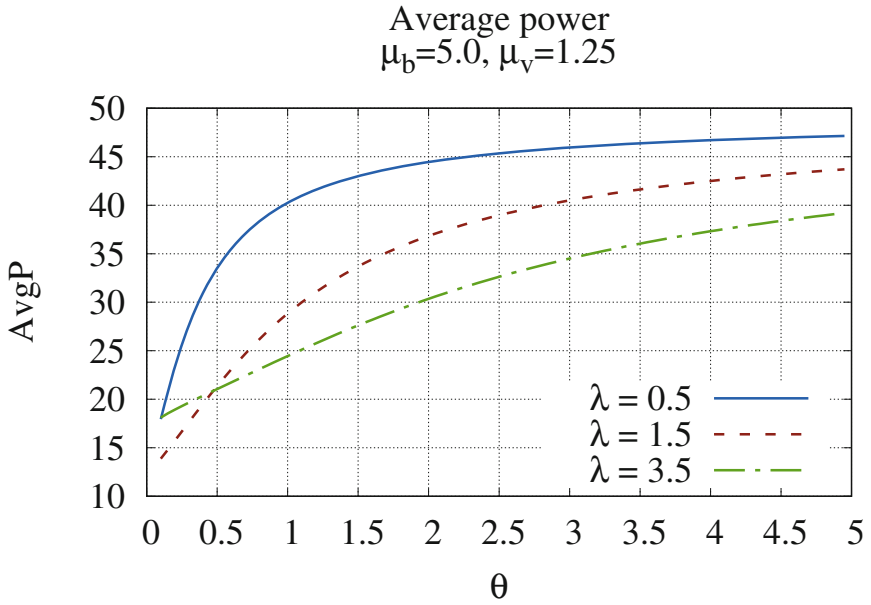


Fig. 11. Average power against  $\vartheta$  in the asynchronous case for  $\lambda \in \{0.5, 1.5, 3.5\}$ .

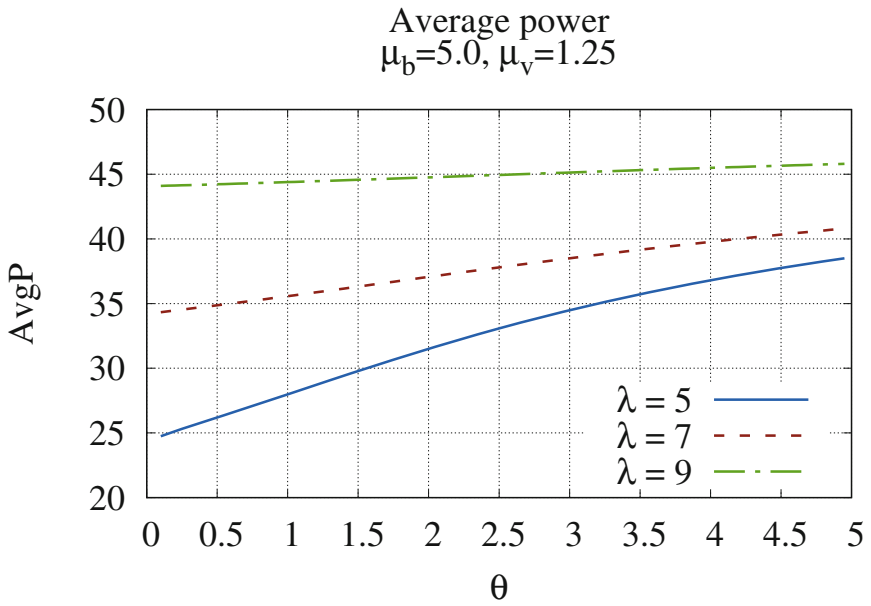


Fig. 12. Average power against  $\vartheta$  in the asynchronous case for  $\lambda \in \{5, 7, 9\}$ .

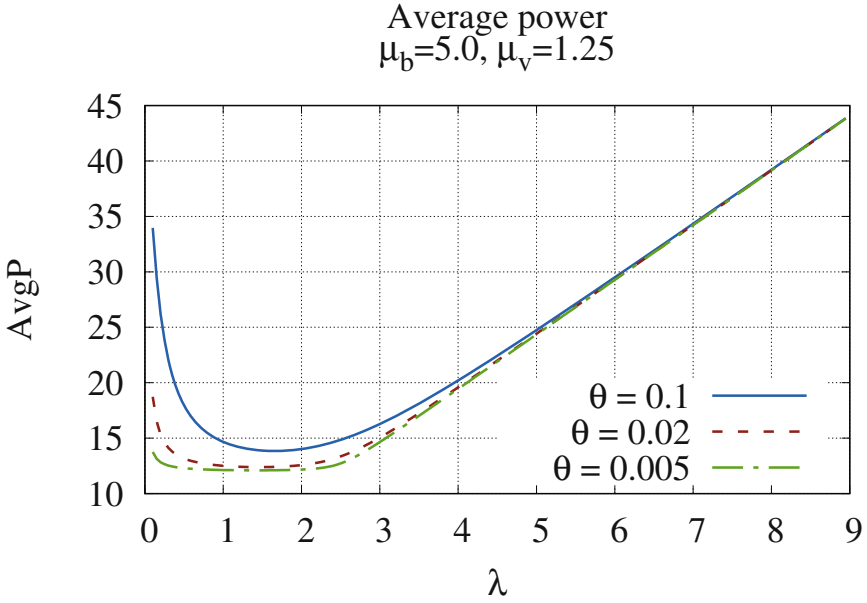


Fig. 13. Average power against  $\lambda$  in the asynchronous case.

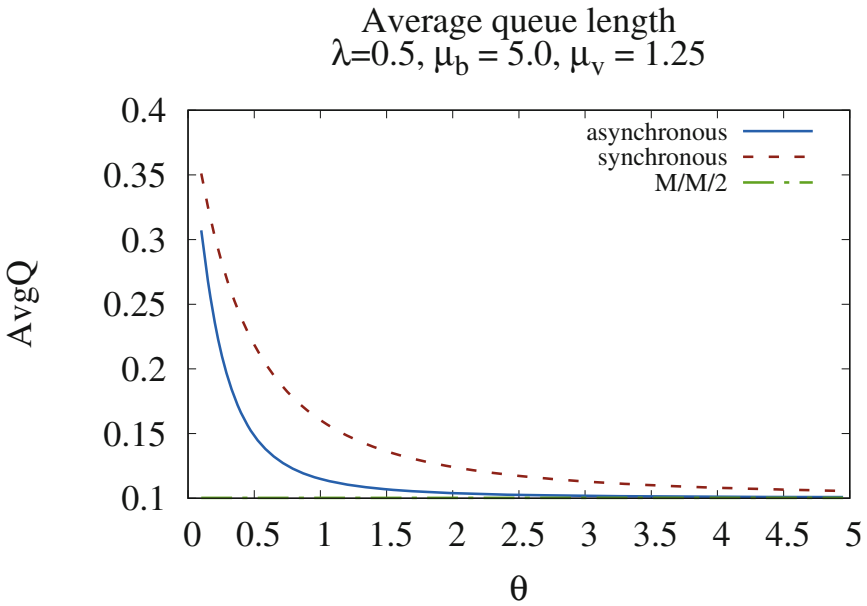


Fig. 14. Average queue length of the synchronous and asynchronous cases for  $\lambda = 0.5$ .

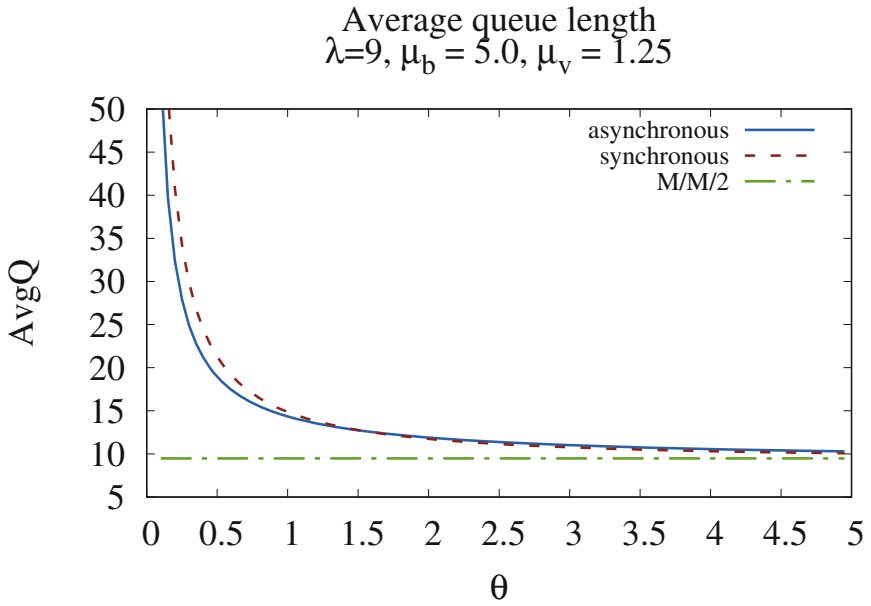


Fig. 15. Average queue length of the synchronous and asynchronous cases for  $\lambda = 9$ .

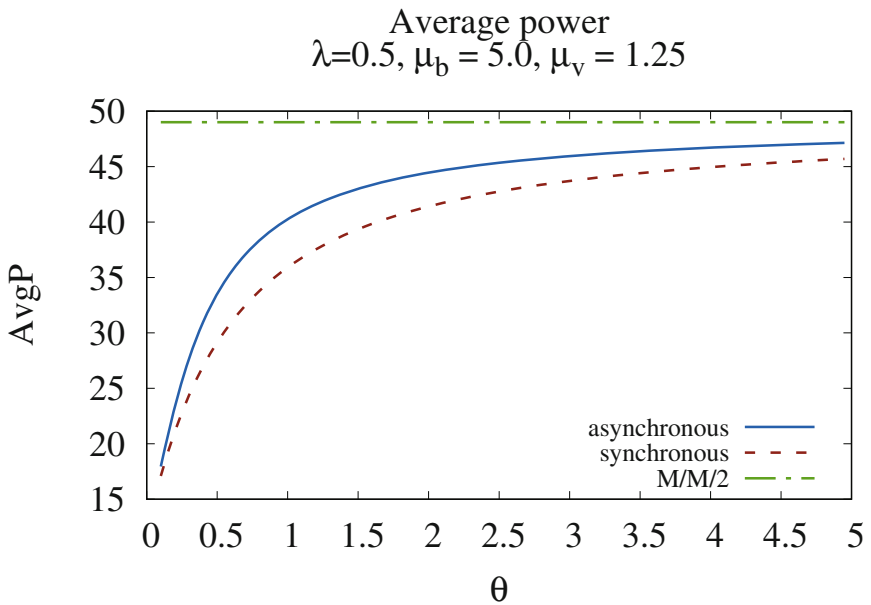


Fig. 16. Average power of the synchronous and asynchronous cases for  $\lambda = 0.5$ .



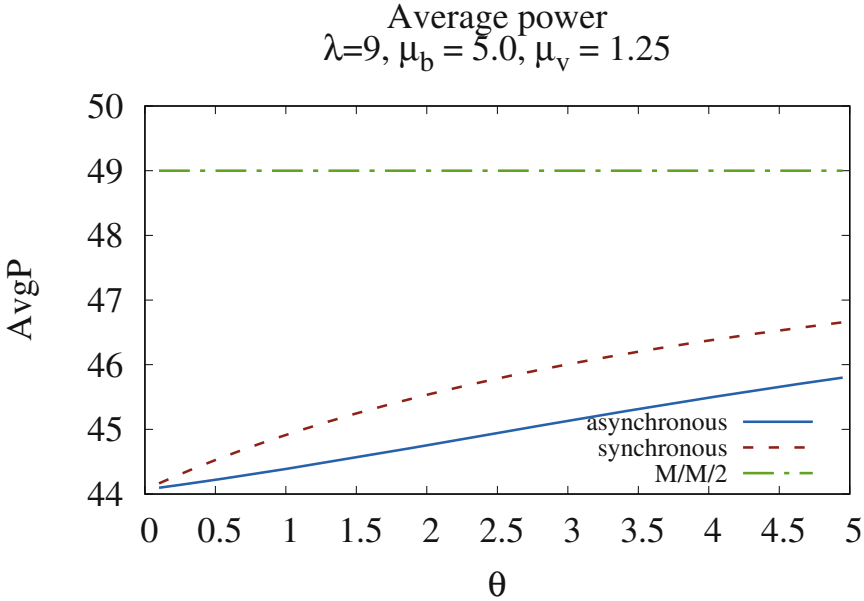


Fig. 17. Average power of the synchronous and asynchronous cases for  $\lambda = 9$ .

asynchronous case. Obviously higher arrival rates ( $\lambda$ ) result in higher average queue lengths. We can also see a steep rise for very low values of  $\vartheta$  due to longer vacations.

Figures 4 and 10 depict the queue length and Figs. 7 and 13 show the average power against the arrival rate  $\lambda$  for  $\vartheta \in \{0.005, 0.02, 0.1\}$  also in the synchronous and the asynchronous cases.

In Figs. 14 and 16 we compared the synchronous, the asynchronous and the simple M/M/2 queue without vacation for  $\lambda = 0.5$ . We did the same in Figs. 15 and 17 but with  $\lambda = 9$ . We can see that for lower  $\lambda$  the asynchronous working vacations produce lower average queue lengths at the cost of higher average power consumption. However at higher  $\lambda$  values the difference is negligible because the high traffic results in fewer vacations in both cases. We should also note that for high traffic the values converge to the case of the simple M/M/2.

## References

1. Altman, E., Yechiali, U.: Analysis of customers' impatience in queues with server vacations. *Queueing Syst.* **52**(4), 261–279 (2006)
2. Baba, Y.: Analysis of a GI/M/1 queue with multiple working vacations. *Oper. Res. Lett.* **33**(2), 201–209 (2005)
3. Chae, K.C., Lim, D.E., Yang, W.S.: The GI/M/1 queue and the GI/Geo/1 queue both with single working vacation. *Perform. Eval.* **66**(7), 356–367 (2009)

4. Lin, C.-H., Ke, J.-C.: Multi-server system with single working vacation. *Appl. Math. Model.* **33**(7), 2967–2977 (2009)
5. Mitrani, I., Chakka, R.: Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. *Perform. Eval.* **23**(3), 241–260 (1995)
6. Servi, L.D., Finn, S.G.: M/M/1 queues with working vacations (M/M/1/WV). *Perform. Eval.* **50**(1), 41–52 (2002)
7. Wang, J., Gao, S., Do, T.V.: Performance analysis of a two-node computing cluster. *Comput. Ind. Eng.* **93**, 227–235 (2016)
8. Zhang, Z.G., Tian, N.: Analysis of queueing systems with synchronous single vacation for some servers. *Queueing Syst.* **45**(2), 161–175 (2003)
9. Zhang, Z.G., Tian, N.: Analysis on queueing systems with synchronous vacations of partial servers. *Perform. Eval.* **52**(4), 269–282 (2003)

# A Review of Technologies for Conversational Systems

Julia Masche<sup>(✉)</sup> and Nguyen-Thinh Le

Department of Informatics, Humboldt-Universität zu Berlin, Berlin, Germany  
julia.masche@hotmail.com,  
nguyen-thinh.le@hu-berlin.de

**Abstract.** During the last 50 years, since the development of ELIZA by Weizenbaum, technologies for developing conversational systems have made a great stride. The number of conversational systems is increasing. Conversational systems emerge almost in every digital device in many application areas. In this paper, we present the review of the development of conversational systems regarding technologies and their special features including language tricks.

## 1 Introduction

Fifty years ago, the chatbot ELIZA was created and considered the first piece of conversational software. The chatbot ELIZA was intended to emulate a psycho-therapist. At that time, it did not pass the Turing test (Turing 1950). Today, conversational computer systems are emerging in many domains, ranging from hotline support over game environments to educational contexts. Some of them can pass the Turing test (e.g., Eugene Goostman (Eugene 2014)). Not only we can find conversational computer systems in many application domains, but smartphones that almost everyone uses daily are integrated with a natural language speech assistant (e.g., “Siri” for iPads, “S-Voice” for *Samsung* tablets/smartphones, “Google Now”), which allows the user to give commands or to ask for information. Recently, “Alexa” speaker of *Amazon* has been developed and is available for English and German speakers. We are facing a change in human-computer interaction: the interaction between humans and computer systems is shifting towards natural language-based interfaces. This paper aims at reviewing the technologies that have been being developed to build conversational systems. Concretely, we investigate the following research questions: Which technologies have been deployed for developing conversational systems? Which language tricks have been commonly exploited? How are typical evaluation methods for conversational systems?

## 2 Methodology

In order to answer the above questions, we searched on the Internet using search machines. Documents that matched the keywords “chatbot”, “conversational agent”, “pedagogical agent”, or “conversational system” were collected. The number of

resulting papers was enormous. Since we intended to investigate technologies for developing conversational systems, we constrained our search based on the following criteria:

1. The conversational system was developed for scientific purposes;
2. The conversational system must have been scientifically evaluated or participated in a competition;
3. Information about the technologies deployed in that system was available.

At the end, we reviewed 59 conversational systems that are summarized in Appendix “Table of reviewed conversational systems”. We categorized the collected systems into “chatbots” and “dialog systems” (Klüwer 2011; Dingli and Scerri 2013; van Woudenberg 2014). The terminology “chatbot” originated from the system CHATTERBOT, which was invented as a game character for the 1989 multiuser dungeon game “TinyMUD” (Mauldin 1994). From the technical point of view, Klüwer (2011) summarized the following typical processing steps of a chatbot: (1) input cleaning (removal and substitution of characters and words like smileys and contractions), (2) using a pattern-matching algorithm to match input templates against the cleaned input, (3) determining the response templates, and (4) generating a response. The second category of conversational systems is “dialog system”. This term denotes a system, which is able to hold a conversation with another agent or with a human. McTear notes the following differences between dialog systems and chatbots: “Dialog systems make use of more theoretically motivated techniques” and “dialog systems often are developed for a specific domain, whereas simulated conversational systems [chatbots] are aimed at open domain conversation.” (McTear 2004). While a typical chatbot is built based on a knowledge base, which comprises a fixed set of input-response templates and a pattern-matching algorithm, a dialogue system typically requires four components: a preprocessing component, a natural language understanding component, a dialog manager, and a response generation component (Lester et al. 2004). The main differences in the architecture between dialog systems and chatbots are the natural language understanding component and the dialog manager.

These two categories of conversational systems are not clearly defined. Rather, these categories describe typical components of each type of conversational systems. A chatbot may also have been implemented using natural language understanding technologies, e.g., LSABot (Agostaro et al. 2005) or overlaps with other components of a typical dialog system. Despite of the overlapping between the two categories, our review is based on them to classify collected conversational systems and their technologies.

## 3 Results

### 3.1 Chatbots

**Pattern Matching.** Pattern matching techniques were used by many chatbots including ELIZA (Weizenbaum 1966), SHRDLU (Winograd 1972; Hutchens 1997),

Speech Chatbot (Senef et al. 1991), PARRY (Colby 1981; Hutchens 1997), PC Therapist III (Weintraub 1986; Hutchens 1997), Chatterbot in “TinyMUD” (Mauldin 1994), TIPS (Whalen 1996; Hutchens 1997), FRED (Garner 1996; Hutchens 1997), CONVERSE (Batacharia et al. 1997; Bradeško and Mladenčić 2012), HEX (Hutchens 1997), Albert One (Garner, 2005; Bradeško and Mladenčić, 2012), Jabberwock (Pirner 2005; Bradeško and Mladenčić 2012). ELIZA, the first chatbot developed by Weizenbaum (1966), deployed pattern matching in order to generate an appropriate response to the user’s utterance. For example, ELIZA would analyze the user’s input “He says I’m depressed much of the time” by matching it to the keywords in a pre-specified dictionary. Then, for a found keyword, ELIZA applies an associated input-response rule. Based on this principle, ELIZA transforms the phrase “I am” into the phrase “You are”. The response generation algorithm adds a phrase “I am sorry to hear” prior to “you are” and a response is generated “I am sorry to hear you are depressed.”

**Cleverscript.** Rollo Carpenter invented the core concepts and developed an algorithm for a chatbot in 1982 (<https://www.existor.com/products/cleverbot-data-for-machine-learning>). In 1996, this algorithm and the chatbot went online under the name “Jabberwacke”. Since 2006, this chatbot was rebranded as Cleverbot and the authoring language Cleverscript for developing chatbots was announced (Cleverscript 2016). The main concept of Cleverscript is based on spreadsheets. Words and phrases that can be recognized (input) or generated (output) by Cleverscript are written on separate lines of the spreadsheet (Jackermeier 2015). Cleverscript and the concept of this chatbot authoring language make the development of chatbots relatively easy. In 2007, Eviebot (<https://www.eviebot.com/en/>), a female embodied chatbot with realistic facial expressions, went online. Additionally, Boibot, a male counterpart for Eviebot, was introduced in 2015 (<https://www.boibot.com/en/>). Both share the same technology with Cleverbot and are able to speak several languages.

**Chatscript.** Chatscript is another authoring language, which serves to facilitate the development of chatbots. Similar to Cleverscript, Chatscript is based on pattern matching (Jackermeier 2015). Another special feature of Chatscript is the so-called Concept Set, which covers semantic-related concepts of a constituent in user input. Chatbots that have been developed using Chatscripts include Suzette (Wilcox and Wilcox 2010), Rosette (Abdul-Kader and Woods 2015), Albert (Latorre-Navarro and Harris 2015), and a conversational agent of Bogatu and colleagues (2015).

**AIML.** In 2001, an XML based language for developing chatbots called AIML was released. The “A.L.I.C.E.” chatbot (Wallace 2003) was the first one developed using this technology. In the past few years, AIML has established itself as one of the most used technologies in today’s chatbots. AIML is based on pattern matching (das Graças Bruno Marietto et al. 2013). An AIML script consists of several “categories”, which are defined by the tag <category>. Each category consists of only one <pattern> tag, which defines a possible user input, and at least one <template> tag, which specifies the chatbot’s response for the user’s input. Like Cleverscript, AIML makes use of wildcards in order to cover a large possibility of user’s inputs. In order to interpret these AIML tags, a chatbot needs an AIML interpreter, which is implemented according to the corresponding AIML specification (either 1.0 or 2.0). Various AIML interpreters

using different programming languages such as Java or Python are available (<http://www.alicebot.org>). Since developing AIML chatbots does not require skills in a specific programming language, this technology facilitates the development of chatbots. Thus, a huge body of chatbots has been developed using AIML technology such as Freudbot (Freudbot 2009), Max (Kopp et al. 2005), the chatbots in (Pilato et al. 2005), Penelope and Alex (Doering et al. 2008), HmoristBot (Augello et al. 2008), chatbot of Alencar et al. (Alencar and Netto 2011), the system of van Rosmalen et al. (2012), Ella (Bradeško and Mladenčić 2012), MathGame (Silververg et al. 2013), Chappie (Behera 2016), and Mitsuku (Abdul-Kader et al. 2015).

**Language Tricks.** In addition to the technologies for chatbots above, we also notice that many chatbots used language tricks in order to fool users and to pass the evaluation. Abdul-Kader (2015) and Bradeško and Mladenčić (2012) summarized four language tricks that are usually used by chatbots including: canned responses, model of personal history, no logical conclusion, typing errors and simulating key strokes. Canned responses are used by chatbots in order to cover questions/answers of the user that are not anticipated in the knowledge based of the chatbot. A model of personal history (e.g., history about the past, childhood stories, social environments, and political and religious attitudes, etc.) enriches the “social background” of a chatbot and pretends the user to a real “person”. Statements with no logical conclusion like “today is today” are embedded in chatbots in order to enrich smalltalks. Typing errors and simulating key strokes are usually used to simulate a “human being” who is typing and making typo errors. HeX (Hutchens 1997), CONVERSE (Batacharia et al. 1997; Bradeško and Mladenčić 2012), PC Therapist III (Bradeško and Mladenčić, 2012), and TIPS (Bradeško and Mladenčić 2012) are conversational systems that make use of one or more language tricks.

### 3.2 Dialog Systems

Based on typical components of a dialog system (Lester et al. 2004), we reviewed the technologies of these components.

**Preprocessing.** Most dialog systems process the user’s input before it is forwarded to the Natural Language Understanding component. The tasks of pre-process are divers. Berger (2014) summarized the following preprocessing tasks of dialog systems: sentence detection, co-resolution, tokenization, lemmatization, POS-tagging, dependency parsing, named entity recognition, semantic role labeling. We found that the dialog systems mostly deployed the following natural language preprocessing tasks: Tokenization (Veselov 2010; Wilks et al. 2010; Eugene 2014; Bogatu et al. 2015; Amilon 2015), POS-Tagging (Lasguido et al. 2013; Dingli et al. 2013; Higashinaka et al. 2014; Ravichandran et al. 2015), sentence detection or chunking (Latorre-Navarro et al. 2015), Named Entity Recognition (Wilks et al. 2010; Lasguido et al. 2013).

**Natural Language Understanding.** The result of preprocessing tasks is ready for the natural language understanding component. For this step, the following approaches are used in dialog systems: Latent Semantic Analysis based on the Vector Space Model

(VSM), e.g. in LSAbot (Agostaro et al. 2005), IRIS (Branchs et al. 2012), AutoTutor (Graesser et al. 1999), Operation ARIES! (Forsyth et al. 2013), dialog system of Pilato et al. (2005); TF-IDF techniques, e.g., Discussion-Bot (Feng et al. 2007).

**Dialog Manager.** The dialogue manager is responsible for coordinating the flow of the conversation in a dialogue system. Approaches to developing dialogue manager are categorized in (1) finite state-based systems, (2) frame-based systems, and (3) agent-based systems Klüwer (2011) and Berger (2014). In finite state-based dialog systems, the flow of the dialogue is specified through a set of dialogue states with transitions denoting various alternative paths through a dialogue graph. At each state, the system produces prompts, recognizes (or rejects) specific words and phrases in response to the prompt, and produces actions based on the recognized response. The dialogue states and their transitions must be designed in advance. Many dialogue systems have been developed applying this approach, e.g. the Nuance automatic banking system (van Woudenberg 2014). Frame-based systems ask the user questions that enable the system to fill slots in a template in order to perform a task such as providing train timetable information. In this type of systems, the dialog flow is not fixed. The dialog flow depends upon the content of the user input, and the information that is elicited by the system. This approach has been used in systems that provide information about movies, train schedules, and the weather. The advantage of the simplicity of these domains is that it is possible to build very robust dialogue systems. One does not need to obtain full linguistic analyses of the user input. The approach underlying agent-based dialog systems is detecting the plans, beliefs and desires of the users and modeling this information in a Belief-Desire-Intention (BDI) agent. Due to the multiple reasoning steps for constructing plans, beliefs and desires of the users, this approach is challenging.

**Response Generation.** The technologies deployed for generating responses are various in different dialog systems. CONVERSE has a generation module, which adds different types of the same expression to an utterance and generates a smooth response (Batacharia et al. 1997). RITEL has a natural language generation module, which is based upon a set of template sentences (Galibert et al. 2005). The proposed conversational system of Higashinaka et al. (2014) combines different modules for utterance generation: the versatile, question answering, personal question answering, topic-inducing, related-word, Twitter, predicate-argument structure, pattern and user predicate-argument structure modules. The generation of utterances applying these modules is based on the last estimated dialogue-act. The conversational agent Albert (Latorre-Navarro et al. 2015) has a language generation module, which consists of templates containing text, pointers, variables and other control functions.

**Special Features.** In addition to technologies for typical dialog systems, we also have learned that conversational systems have been implemented with special features in order to make them more likely “humans”. For instance, some systems are able to learn from conversations and can apply this knowledge later. The chatbot MegaHal (Hutchens 1997; Hutchens et al. 1998) talks a lot of gibberish in order to fool its user, whereas the system Ella (Bradeško and Mladenić 2012) is able to spot gibberish initiated by the user and react in an appropriate way. Moreover, there are many multimodal systems (Ferguson et al. 1996; Bickmore et al. 2000; Bohus et al. 2004; Pradhan et al. 2016), which can communicate with the user through both text and

speech channels. With the development of embodied conversational agents, features like gestures, facial expressions or eye gazes become increasingly important (Alexander et al. 2006; Ayedoun et al. 2015). Developers of pedagogical agents also often include graphics, videos, animations and interactive simulations into their system to increase the student's motivation (Kim et al. 2007; Forsyth et al. 2013; Pradhan et al. 2016).

### 3.3 Evaluation Methods

Since we only collected conversational systems that have been evaluated or participated in a competition contest, we categorized the evaluation methods that have been used into four classes: (1) qualitative analysis, (2) quantitative analysis, (3) pre-/posttest, and (4) chatbot competitions. Note, that many systems may have been evaluated using more than just one evaluation method.

The first most applied evaluation method was the quantitative method, which used interviews or questionnaires. Examples of conversational systems that have been evaluated using this method include, e.g., Speech Chatbot (Senef et al. 1991), TRAINS-95 (Ferguson et al. 1996; Sikorski and Allen 1996), Herman the Bug in Design-A-Plant (Lester et al. 1997), REA (Bickmore et al., Bickmore and Cassell 2000), LARRI (Bohus et al. 2004), FAQchat (Shawar et al. 2005), Discussion-Bot (Feng et al. 2007), Freudbot (Freudbot 2009), Justin and Justina (Kenny et al. 2011), the dialogue system of Shibata et al. (2014), or Pharmabot (Comendador et al. 2015).

The second widely used evaluation method is quantitative. The quantitative method makes use of dialog protocols generated by conversations between the user and the system. Examples of conversational systems that have been evaluated using this method include RAILTEL (Bennacef et al. 1996), Max (Kopp et al. 2005), HumoristBot (Augello et al. 2008), Senior Companion (Wilks et al. 2010, 2008), SimStudent (MacLellan et al. 2014), Betty's Brain (Leelawong et al. 2008; Biswas et al. 2005), CALMsystem (Kerly et al. 2007), Discussion-Bot (Feng et al. 2007), the dialogue system of Planells et al. (2013), or Albert (Latorre-Navarro et al. 2015).

The third evaluation method deploys pre- and post-tests. The method has been used usually for evaluating pedagogical agents to measure the learning effect. This method was applied for the evaluation of MathGirls (Kim et al. 2007), My Science Tutor (Pradhan et al. 2016), Herman the Bug (Lester et al. 1997) or MetaTutor (Bouchet et al. 2013; Harley et al. 2014).

The fourth evaluation method is the participation of a conversational system in a competition contest, for example, the Loener prize, which is based on the Turing Test (Abdul-Kader et al. 2015). Loebner Prize winners were, for instance, PARRY (Colby 1981; Hutchens 1977), CONVERSE (Batacharia et al. 1997; Bradevsko et al. 2012), A. L.I.C.E (Wallace 2003), Albert One (Garner 2005; Bradeško and Mladenčić 2012), Elbot (Abdul-Kader et al. 2015), and Mitsuku (Abdul-Kader et al. 2015).



## 4 Discussion and Conclusions

In this paper, we have reviewed the technologies, language tricks, special features, and evaluation methods of conversational systems. While chatbots deploy dominantly pattern matching techniques and language tricks, most dialog systems exploit natural language technologies. We also have learned that most chatbots participated in the Turing test contests (e.g., Loebner prize), while dialog systems were mostly evaluated by the pre-/post-test, quantitative, or qualitative methods. This can be explained by the fact that dialog systems are more goal-oriented (e.g., to improve learning gains of students) and chatbots rather serve smalltalks in different domains. Based on the summary table in Appendix, we can notice the tendency of applied technologies for conversational systems: they are becoming more AI-oriented and deploying more natural language processing technologies.

In this paper, due to the page limit, we summarized the technologies for developing conversational systems. We plan to elaborate on these technologies in more details in a journal article.

### Appendix: Table of Reviewed Conversational Systems

Year	Category (Name)	Technology
1966	Chatbot (ELIZA) (Weizenbaum 1966)	Pattern matching; keyword searching
1971	Chatbot (PARRY) (Colby 1981; Hutchens 1997)	Parsing, interpretation-action-module
1972	Chatbot (SHRDLU) (Winograd 1972; Hutchens 1997)	Parsing, grammatical detection, semantic analysis
1991	Chatbot (PC Therapist III) (Bradeško and Mladenić 2012)	Parsing, pattern matching, knowledge database (quotes & phrases)
1991	Speech Chatbot (Senef et al. 1991)	Parsing, Response Generator, Semantic Frame Representation, Pattern Matching
1994	Chatbot (TIPS) (Bradeško and Mladenić 2012)	Pattern matching, system similar to a database
1994	Chatbot (Chatterbot in TinyMud) (Mauldin 1994)	Pattern matching, Markov chain models
1996	Chatbot (HeX) (Hutchens 1997)	Pattern matching, Markov chain models
1996	Chatbot (Jabberwacky/Cleverbot) (Cleverbot 2016)	Cleverscript
1996	Speech Dialog System (TRAINS-95) (Ferguson et al. 1996)	Bottom-up chart parser; Discourse manager; Text generator; Language understanding; Verbal reasoner
1996	Speech Dialog System (RAILTEL) (Bennacef et al. 1996)	Speech recognition; Literal and contextual understanding; Parser, Dialog manager

(continued)

*(continued)*

Year	Category (Name)	Technology
1997	Chatbot (CONVERSE) (Batacharia et al. 1997; Bradeško and Mladenčić, 2012)	Input module; Pre-processing; Parser; Pattern matching; WordNet synonyms; ontology; fact & person database; Action module; Topic change module; Utterance generator
1997	Speech Dialog System (Herman the Bug) (Lester et al. 1997)	Coherence-Structured Behavior Space Framework, Behaviour Control
1998	Chatbot (MegaHal) (Hutchens 1998)	Markov chain models; Keyword matching
1999	Dialog system (AutoTutor) (Graesser et al. 1999)	NLP (POS tags); Dialog move generator; Latent semantic analysis; Regular expression matching; Speech act classifiers
1998–99	Chatbot (Albert One) (Garner 2005; Bradeško and Mladenčić 2012)	Pattern matching
2000	Speech Dialog System (REA) (Bickmore et al. 2000)	Discourse planner; Natural language generation engine
2000/01	Chatbot (A.L.I.C.E) (Wallace 2003)	AIML
2001	Chatbot (Eugene Goostman) (Eugene 2014; Veselov 2010)	Advanced Pattern Matching; Tokenization (dynamic)
2002	Chatbot (Ella) (Bradeško and Mladenčić 2012)	Pattern matching, AIML, WorldNet
2003	Chatbot (Jabberwock) (Pirner 2005; Bradeško and Mladenčić, 2012)	Parsing (Context Free Grammar); Pattern Matching; Markov Chains Models
2004	Speech Dialog System (LARRI) (Bohus et al. 2004)	Speech recognition; Dialog manager; Response generator; Parsing with semantic Grammar; Task Markup Language
2005	Dialog System (Freudbot) (Freudbot 2009)	AIML
2005	Dialog System (LSAbot) (Agostaro et al. 2005)	LSA, AIML (same knowledge database as ALICE)
2005	Dialog System (FAQchat) (Shawar et al. 2005)	Advanced Pattern Matching
2005	Embodied Dialog System (Max) (Kopp et al. 2005)	NLP, Dialog Manager; Interpreter; AIML
2005	Speech Dialog System (RITEL) (Galibert et al. 2005; Toney et al. 2008)	Speech Recognition; Parsing; Input Analysis, Named Entity Analysis; Lexical Analysis; Dialog Manager; Response Generator
2005	Chatbot (Pilato et al. 2005)	AIML, Latent Semantic Analysis

*(continued)*

(continued)

Year	Category (Name)	Technology
2006	Speech Dialog System (Eve) (Alexander et al. 2006; Sarrafzadeh et al. 2014)	Facial expression analysis; Case-Based Methods; Dialog Manager
2007	Dialog System (CALMsystem) (Kerly et al. 2007)	NLP techniques; Pattern Matching
2007	Chatbot (Discussion-Bot) (Feng et al. 2007)	Information-Retrieval and NLP techniques
2007	Speech Dialog System (MathGirls) (Kim et al. 2007)	Relational database where actions are stored
2008	Dialogue System (Penelope and Alex) (Doering et al. 2008)	AIML
2008	Dialogue System (Betty's Brain) (Leelawong et al. 2008; Biswas et al. 2005)	Qualitative Reasoning Methods; Perception System; Knowledge Database
2008	Dialog System (HumoristBot) (Augello et al. 2008)	AIML
2010	Chatbot (Suzette) (Wilcox and Wilcox 2010)	ChatSript
2010	Dialog System (Senior Companion) (Wilks et al. 2010; Wilks et al. 2008)	Tokenization; POS tagging; Parsing; Information Extraction techniques (Named Entity Recognition); Reasoner; Dialog Manager
2011	Chatbot (Rosette) (Abdul-Kader et al. 2015)	ChatScript
2011	Speech Embodied Dialog System (JUSTIN und JUSTINA) (Kenny et al. 2011)	Speech Recognition; Parsing, Question-Answering; Pattern Matching; Dialog Manager; Response Generator
2011	Chatbot (Alencar et al. 2011)	AIML
2011	Dialog System (Operation ARIES!) (Forsyth et al. 2013)	Training Module; LSA; Regular Expressions; Pattern Matching
2012	Chatbot (IRIS) (Branchs et al. 2012)	Vector Space Model
2012	Chatbot (van Rosmalen et al. 2012)	Lexical analysis, AIML, semantic structure
2013	Chatbot (Mitsuku) (Abdul-Kader et al. 2015)	AIML
2013	Dialog System (Lasguido et al. 2013)	Dialog Manager; POS-Tagging, NER, Similarity Search
2013	Dialog System (Math Game) (Silvervarg et al. 2013)	AIML; Dialog Manager
2013	Dialog System (MetaTutor) (Bouchet et al. 2013; Harley et al. 2014)	NLP, Dialog Manager, Parsing, XML, Facial Expression Analysis

(continued)

*(continued)*

Year	Category (Name)	Technology
2013	Speech Dialog System (Planells et al. 2013)	Speech recognition; Task Manager, Multimodal Response Generator
2013	Dialog System (Prototype) (Dingli et al. 2013)	ChatScript, POS-Tagging, Response Generator
2014	Dialog System (Higashinaka et al. 2014)	Natural language understanding (Noun Phrase removal, POS-Tagging); Dialog Manager; Response Generator; Question-Answering; Pattern Matching
2014	Dialog System (Shibata et al. 2014)	Dialog Manager, Response Generator, Markov Chain Models
2014	Dialog System (SimStudent) (MacLellan et al. 2014)	Machine-Learning techniques; Regular Expressions
2014	Dialog System (My Science Tutor, Virtual Tutor Marni) (Pradhan et al. 2016)	Parsing; Semantic annotation
2015	Chatbot (Amilon 2015)	Parsing (Parser Trees); ConceptNet
2015	Speech Dialog System (dBot) (Ravichandran et al. 2015)	parsing, POS-Tagging, Noun-Phrase Extraction + Artificial Intelligence algorithm
2015	Dialog System (Model of Historical Personality) (Bogatu et al. 2015)	ChatScript, POS-Tagging, Tokenization, Lemmatization; Morphological analysis
2015	Dialog System (Pharmabot) (Comendador et al. 2015)	Parsing
2015	Dialog System (Albert) (Latorre-Navarro et al. 2015)	Dialog manager; ChatScript; NLU (POS Tags; Noun-Phrase Chunking; Lexical Relations)
2016	Chatbot (Chappie) (2016)	NLP; AIML; Response Generator

## References

- Abdul-Kader, S.A., Woods, J.: Survey on Chatbot design techniques in speech conversation systems. *Int. J. Adv. Comput. Sci. Appl.* **6**(7), 72–80 (2015)
- Agostaro, F., Augello, A., Pilato, G., Vassallo, G., Gaglio, S.: A conversational agent based on a conceptual interpretation of a data driven semantic space. In: *Advances in Artificial Intelligence*. LNCS, vol. 3673, pp. 381–392 (2005)
- Alexander, S., Sarrafzadeh, A., Hill, S.: Easy with eve: a functional affective tutoring system. In: Rebolledo-Mendez, G., Martinez-Miron, E. (eds.) *Workshop on Motivational and Affective Issues in ITS* held at the 8th International Conference on ITS, pp. 38–45 (2006)
- Alencar, M., Netto, J.M.: In *Proceedings - Frontiers in Education Conference* (2011)
- Amilon, M.: Chatbot with common-sense database. Bachelor's thesis in Computer Science, KTH Royal Institute of Technology, Sweden (2015)

- Augello, A., Saccone, G., Gaglio, S., Pilato, G.: Humorist bot: bringing computational humour in a Chat-Bot system. In: Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems (2008)
- Ayedoun, E., Hayashi, Y., Seta, K.: A conversational agent to encourage willingness to communicate in the context of english as a foreign language. *J. Procedia Comput. Sci.* **60**, 1433–1442 (2015). ScienceDirect
- Batacharia, B., Levy, D., Catizone, R., Krotov, A., Wilks, Y.: CONVERSE: a conversational companion. In: Wilks, Y. (ed.) *Machine Conversations. The Springer International Series in Engineering and Computer Science*, vol. 511. Springer, Heidelberg (1997)
- Behera, B.: Chappie-A Semi-automatic Intelligent Chatbot (2016). [https://www.cse.iitb.ac.in/~bibek/WriteUP\\_2016.pdf](https://www.cse.iitb.ac.in/~bibek/WriteUP_2016.pdf). Accessed 02 May 2017
- Bennacef, S., Devillers, L., Rosset, S., Lamel, L.: *Dialog in the RAILTEL Telephone-Based System* (1996)
- Berger, M.: *Modelling of Natural Dialogues in the Context of Speech-based Information and Control Systems*. Ph.D. dissertation submitted to Christian-Albrechts-Universität zu Kiel. AKA. IOS Press (2014)
- Bickmore, T., Cassell, J.: “How about this weather?” Social Dialogue with Embodied Conversational Agents. American Association for Artificial Intelligence (2000)
- Biswas, G., Schwartz, D., Leelawong, K., Vye, N., TAG-V: Learning by teaching a new agent paradigm for educational software. *Int. J. Appl. Artif. Intell.* **19**(3–4) (2005)
- Bogatu, A., Rotarescu, T., Rebedea, T., Ruseti, S.: *Conversational Agent that Models a Historical Personality* (2015)
- Bohus, D., Rudnicky, A.I.: LARRI: a language-based maintenance and repair assistant. Computer Science Department (2004). <http://repository.cmu.edu/compsci/1347>
- Bouchet, F., Harley, J.M., Azevedo, R.: Impact of different pedagogical agents’ adaptive self-regulated prompting strategies on learning with MetaTutor. In: Proceedings of the International Conference on Artificial Intelligence in Education, pp. 815–819 (2013)
- Bradeško, L., Mladenčić, D.: A survey of chatbot systems through a Loebner prize competition. Artificial Intelligence laboratory, Jozef Stefan Institute, Ljubljana Slovenia (2012)
- Branchs, R.F., Li, H.: IRIS: a chat-oriented dialogue system based on the vector space model. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 37–42. Association for Computational Linguistics (2012)
- Cleverbot: Data for Machine Learning. Conversation Data API on (2016). <https://www.existor.com/products/cleverbot-data-for-machine-learning/>. Accessed 12 Feb 2017
- Cleverscript (2016). <https://www.existor.com/products/cleverscript/>. Accessed 6 Dec 2017
- Colby, K.M.: Modeling a paranoid mind. *Behav. Brain Sci.* **4**(4), 515–534 (1981). doi:10.1017/S0140525X00000030. Cambridge University Press
- Comendador, B.E.V., Francisco, B.M.B., Medenilla, J.F., Nacion, S.M.T., Serac, T.B.E.: Pharmabot: a pediatric generic medicine consultant Chatbot. *J. Autom. Control Eng.* **3**(2), 137–140 (2015)
- das Graças Bruno Marietto, M., de Aguiar, R.V., de Oliveira Barbosa, G., Botelho, W.T., Pimentel, E., dos Santos Franca, R., da Silva, V.L.: Artificial intelligence markup language: a brief tutorial (2013)
- Dingli, A., Scerri, D.: Dialog systems and their inputs. In: HCI International 2013 - Posters’ Extended Abstracts. Communications in Computer and Information Science, vol. 374, pp. 601–605 (2013)
- Doering, A., Veletsianos, G., Yerasimou, T.: Conversational agents and their longitudinal affordances on communication and interaction. *J. Interact. Learn. Res.* **19**(2), 251–270 (2008). AACE

- Eugene: Turing Test Success Marks Milestone in Computing History (2014). Accessed 08 Dec 2016
- Feng, D., Shaw, E., Kim, J., Hovy, E.: An intelligent discussion-bot for answering student queries in threaded discussions. Interaction Challenges for Intelligent Assistants, Papers from the 2007 AAAI Spring Symposium, Technical report SS-07-04 (2007)
- Ferguson, G., Allen, J., Miller, B.: TRAINS-95: towards a mixed-initiative planning assistant. In: Proceedings of the 3rd Conference on Artificial Intelligence Planning Systems, pp. 70–77. AAAI (1996)
- Forsyth, C.M., Graesser, A.C., Pavlik Jr., P., Cai, Z., Butler, H., Halpern, D., Millis, K.: Operation ARIES!: methods, mystery, and mixed models: discourse features predict affect in a serious game. *J. Educ. Data Min.* **5**(1), 147–189 (2013)
- Freudbot (2009). <https://fhss2.athabascau.ca/html/Freudbot/test.html>. Accessed 12 Feb 2017
- Galibert, O., Illouz, G., Rosset, S.: Ritel: an open-domain, human-computer dialog system. In: Proceedings of the 9th European Conference on Speech Communication and Technology (2005)
- Garner, R.: The idea of FRED. *ALMA-Scores of the Unfinished Thought*, Issue 1 (1996)
- Garner, R.: Multifaceted Conversational Systems. Colloquium on Conversational Systems, University of Surrey (2005). <http://www.robtron.com/Robby/Multifaceted.ppt>. Accessed 04 May 2017
- Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R.: AutoTutor: a simulation of a human tutor. *J. Cogn. Syst. Res.* **1**, 35–51 (1999). Oden, G.C. (ed.)
- Harley, J.M., Bouchet, F., Papaioannou, N., Carter, C., Trevors, G., Feyzi-Beghnagh, R., Azevedo, R., Landis, R.S.: Assessing learning with MetaTutor, a multi-agent hypermedia learning environment. In: Symposium on Innovative Practices for Assessing Learning in Computer Based Learning Environments, American Educational Research Association (2014)
- Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing. In: Proceedings of the 25th International Conference on Computational Linguistics, pp. 928–939 (2014)
- Hutchens, J.L.: How to Pass the Turing Test by Cheating (1997)
- Hutchens, J.L., Alder, M.D.: Introducing MegaHal. In: Proceedings of the Joint Conference on New Methods in Language Processing and Computational Natural Language Learning, pp. 271–274 (1998)
- Jackermeier, R.: Analyse von Chatbot-Beschreibungssprachen AIML 2.0 im Vergleich zu ChatScript und Cleverscript (2015)
- Kenny, P.G., Parsons, T.D.: Embodied conversational virtual patients. In: Perez-Marin, D., Pascual-Nieto, I. (eds.) *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices* (2011)
- Kerly, A., Ellis, R., Bull, S.: CALMsystem: a conversational agent for learner modelling. In: Proceedings of AI-2007, 27th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, pp. 89–102. Springer (2007)
- Kim, Y., Wei, Q., Xu, B., Ko, Y., Ilieva, V.: MathGirls: toward developing girls' positive attitude and self-efficacy through pedagogical agents. In: Luckin, R., Koedinger, K.R., Greer, J. (eds.) *Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, vol. 158, pp. 119–126. IOS Press (2007)
- Klüwer, T.: From Chatbots to Dialog Systems. In: Perez-Marin, D., Pascual-Nieto, I. (eds.) *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices* (2011)
- Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application (2005)

- Lasguido, N., Sakti, S., Neubig, G., Toda, T., Adriani, M., Nakamura S.: Developing non-goal dialog system based on examples of drama television. In: *Natural Interaction with Robots, Knowbots and Smartphones*, pp. 355–361 (2013)
- Latorre-Navarro, E.M., Harris, J.G.: An intelligent natural language conversational system for academic advising. (IJACSA) *Int. J. Adv. Comput. Sci. Appl.* **6**(1), 110–119 (2015)
- Leelawong, K., Biswas, G.: Designing learning by teaching agents: the Betty’s brain system. *Int. J. Artif. Intell. Educ.* **18**(3), 181–208 (2008)
- Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., Bhogal, R.S.: The persona effect: affective impact of animated pedagogical agents. In: *Proceedings of CHI*, pp. 359–366. ACM Press (1997)
- Lester, J., Branting, K., Mott, B.: Conversational agents. In: Singh, M.P. (ed.) *The Practical Handbook of Internet Computing*. Chapman & Hall/CRC, Boca Raton (2004)
- MacLellan, C.J., Wiese, E.S., Matsuda, N., Koedinger, K.R.: *SimStudent: Authoring Expert Models by Tutoring* (2014)
- Mauldin, M.L.: Chatterbots, TinyMuds, and the turing test: entering the Loebner Prize competition. In: *Proceedings of the 12th National Conference on Artificial Intelligence*, vol. 1, pp. 16–21. American Association for Artificial Intelligence (1994)
- McTear, M.F.: *Spoken Dialogue Technology-Toward the Conversational User Interface*. Springer, London (2004)
- Pilato, G., Vassallo, G., Augello, A., Vasile, M., Gaglio, S.: Expert chat-bots for cultural heritage. *Intelligenza Artificiale* **2**, 25–31 (2005)
- Pirner, J.: The beast can talk (2005). <http://www.abenteuermedien.de/jabberwock/how-jabberwock-works.pdf>. Accessed 04 May 2017
- Planells, J., Hurtado, L., Segarra, E., Sanchis, E.: A multi-domain dialog system to integrate heterogeneous SpokenDialog systems. In: *Proceedings of 14th Annual Conference of the International Speech Communication Association* (2013)
- Pradhan, S., Cole, R., Ward, W.: My science tutor: learning science with a conversational virtual tutor. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics - System Demonstrations*, pp. 121–126 (2016)
- Ravichandran, G., Reddy, N.M., Shriya G.S.: dBot: AI based conversational agent. *Int. J. Sci. Res. (IJSR)* (2015). ISSN (Online): 2319-7064, 1277-1279
- Sarrafzadeh, A., Fourie, L., Kingston, T., Alexander, S.T.V., Overmyer, S., Shanbehzadeh, J.: Eve: an affect-sensitive pedagogical agent. In: *Global Business and Technology Association Conference*, pp. 573–579 (2014)
- Senef, S., Glass, J.G., Hirschmann, L., Polifroni, J.: Development and preliminary evaluation of the MIT ATIS system. In: *Proceedings of the Workshop on Speech and Natural Language*, pp. 88–93. Association for Computational Linguistics (1991)
- Shawar, B.A., Atwell, E., Roberts, A.: FAQchat as an information retrieval system. In: *Proceedings of the 2nd Language and Technology Conference Human Language Technologies as a Challenge*, pp. 274–278 (2005)
- Shibata, T., Egashira, Y., Kurohashi, S.: Chat-like conversational system based on selection of reply generating module with reinforcement learning. In: *Proceedings of 5th International Workshop on Spoken Dialog Systems* (2014)
- Sikorski, T., Allen, J.F.: *TRAINS-95 System Evaluation* (1996)
- Silvervarg, A., Jönsson, A.: Towards a conversational pedagogical agent capable of affecting attitudes and self-efficacy (2013)
- Toney, D., Rosset, S., Max, A., Galibert, O., Bilinski, E.: An evaluation of spoken and textual interaction in the RITEL interactive question answering system. In: *Proceedings of the International Conference on Language Resources and Evaluation* (2008)
- Turing, A.M.: *Computing Machinery and Intelligence*. *Mind* **49**, 433–460 (1950)

- Van Rosmalen, P., Eikelboom, J., Bloemers, E., van Winzum, K., Spronck, P.: Towards a game-chatbot: extending the interaction in serious games. In: Proceedings of the 6th European Conference on Games Based Learning (2012)
- Van Woudenberg, A.F.: A Chatbot dialogue manager, Chatbots and dialogue systems: a hybrid approach. Master thesis, Open University of the Netherlands (2014)
- Veselov, V.: Eugene Goostman (2010)
- Wallace, R.S.: The elements of AIML style. In: 2003 ALICE A.I. Foundation, Inc. (2003)
- Weintraub, J.: History of the PC Therapist (1986). <http://www.loebner.net/Prizef/weintraub-bio.html>.
- Weizenbaum, J.: A computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**(1), 36–45 (1966)
- Whalen, T.: My experience at Loebner prize, ALMA, Issue 1 (1996)
- Wilcox, B., Wilcox, S.: Suzette, the Most Human Computer (2010)
- Wilks, Y., Worgan, S.: A prototype for a Conversational Companion for reminiscing about images. In: *Computer Speech & Language* (2008). doi:[10.1016/j.csl.2010.04.002](https://doi.org/10.1016/j.csl.2010.04.002)
- Wilks, Y., Catzone, R., Dingli, A., Cheng, W.: Demonstration of a prototype for a Conversational Companion for reminiscing about images. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (2010)
- Winograd, T.: *Understanding Natural Language*. Academic Press, Cambridge (1972)



# Author Index

## A

Aguirre, Carlos, [84](#)  
Apolinario, Oscar, [141](#)  
Araya, Roberto, [84](#)  
Ashida, Atsushi, [107](#)

## B

Bajaria, Pratik, [60](#)  
Baklanov, Artem, [47](#)  
Bhopale, Prashant, [60](#)  
Boonkwan, Prachya, [184](#)  
Bradford, Roger, [153](#)

## C

Calfucura, Patricio, [84](#)  
Coenen, Jarno, [72](#)

## D

del Pilar Salas-Zárate, María, [141](#)  
Della Ventura, Michele, [165](#)  
Do, T.V., [197](#)

## F

Farrenkopf, Tobias, [175](#)

## G

Gross, Sebastian, [72](#)

## J

Jaure, Paulina, [84](#)

## K

Kazi, Faruk, [60](#)  
Kojiri, Tomoko, [107](#)  
Krumrey, Linus, [34](#)

## L

Lagos-Ortiz, Katty, [141](#)  
Le Thi, Hoai An, [1](#)

Le, Hoai Minh, [1](#)  
Le, Nguyen-Thinh, [212](#)  
Le, Nguyen-Truong, [175](#)  
Luna-Aveiga, Harry, [141](#)

## M

Masche, Julia, [212](#)  
McCollum, David, [47](#)  
Medina-Moreira, José, [141](#)  
Moeini, Mahdi, [34](#)

## N

Nguyen, D.V., [13](#)  
Nguyen, Dinh, [129](#)  
Nguyen, Hai T., [197](#)  
Nguyen, Linh T.T., [117](#)  
Nguyen, Loan T.T., [117](#)  
Nguyen, Mao, [129](#)  
Nguyen, Thi Hai Binh, [95](#)

## P

Paredes-Valverde, Mario Andrés, [141](#)  
Pham, Dang, [129](#)  
Phan, Duy Nhat, [1](#)  
Pinkwart, Niels, [72](#)

## Q

Quan, Tho, [129](#)

## S

Singh, Navdeep, [60](#)  
Son, Dang Vu, [23](#)  
Subkhankulova, Dina, [47](#)  
Supnithi, Thepchai, [184](#)

## T

Tran, Bach, [1](#)  
Tran, Phong Nha, [95](#)  
Tran, Quang Dieu, [95](#)

Tran, Quang Hai Bang, [95](#)

Tran, Thi Thanh Nga, [95](#)

Tuan, Nguyen Nhu, [23](#)

## V

Valencia-García, Rafael, [141](#)

Vo, Bay, [117](#)

Vo, Khuong, [129](#)

## W

Warschat, Joachim, [175](#)

Wendt, Oliver, [34](#)