IFIP AICT 456

Abdelmalek Amine
Ladjel Bellatreche
Zakaria Elberrichi
Erich J. Neuhold
Robert Wrembel
(Eds.)

Computer Science and Its Applications

5th IFIP TC 5 International Conference, CIIA 2015
Saida, Algeria, May 20–21, 2015
Proceedings

Springer

IFIP Advances in Information and Communication Technology 456

More information about this series at http://www.springer.com/series/6102

Abdelmalek Amine · Ladjel Bellatreche
Zakaria Elberrichi · Erich J. Neuhold
Robert Wrembel (Eds.)

Computer Science and Its Applications

5th IFIP TC 5 International Conference, CIIA 2015
Saida, Algeria, May 20–21, 2015
Proceedings

Springer

Editors
Abdelmalek Amine
Tahar Moulay University
Saida
Algeria

Ladjel Bellatreche
LIAS/ISAE-ENSMA
Chasseneuil
France

Zakaria Elberrichi
Sidi Bel Abbès University
Sidi Bel Abbès
Algeria

Erich J. Neuhold
University of Vienna
Vienna
Austria

Robert Wrembel
Poznan University of Technology
Poznan
Poland

ISSN 1868-4238 ISSN 1868-422X (electronic)
IFIP Advances in Information and Communication Technology
ISBN 978-3-319-19577-3 ISBN 978-3-319-19578-0 (eBook)
DOI 10.1007/978-3-319-19578-0

Library of Congress Control Number: 2015940015

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains research papers presented at the 5th IFIP International Conference on Computer Science and Its Applications (CIIA), held during May 20-21, 2015, in Saida, Algeria. CIIA 2015 continued the series of conferences whose main objective is to provide a forum for the dissemination of research accomplishments and to promote the interaction and collaboration between various research communities related to computer science and its applications. These conferences have been initiated by researchers from Algeria and extended to cover worldwide researchers focusing on promoting research, creating scientific networks, developing projects, as well as facilitating faculty and student exchanges, especially in Africa.

This year the CIIA conference attracted 225 submissions from 20 countries including: Algeria, Bangladesh, Belgium, Canada, China, Finland, France, India, Iran, Ireland, Italy, Jordan, Morocco, Norway, Poland, Qatar, Tunisia, United Arab Emirates, UK, and USA. In a rigorous reviewing process, the Program Committee (PC) selected 51 papers, which represents an acceptance rate of 22.6%. The PC included 200 researchers from 27 countries.

The accepted papers were organized into the four following research tracks: Computational Intelligence, co-chaired by: Sadok Ben Yahia (FST, Tunisia) and Nadjet Kamel (Setif University, Algeria); Security and Network Technologies, co-chaired by Nadjib Badache (CERIST, Algeria) and Alfredo Cuzzocrea (ICAR-CNR and University of Calabria, Italy); Information Technology, co-chaired by Jorge Bernardino (ISEC-Polytechnic Institute of Coimbra, Portugal) and Selma Khouri (ESI, Algiers, Algeria); as well as Software Engineering, co-chaired by Kamel Barkaoui (CNAM, Paris, France) and Abdelwahab Hamou-Lhadj (Concordia University, Montreal, Canada). Additionally, the conference hosted three keynote speakers, namely: Prof. Lynda Tamine-Lechani (IRIT Toukouse, France), Prof. Erich Neuhold (University of Vienna, Austria), and Prof. Mohamad Sawan (Polytechnique Montreal, Canada). This volume includes the abstracts of the keynote talks. We would like to express our warmest thanks to the keynote speakers.

We would also like to extend our gratitude to Prof. Erich Neuhold and the International Federation for Information Processing (IFIP) for accepting the CIIA papers to be published in the *IFIP Advances in Information and Communication Technology* (IFIP-AICT) by Springer.

We would also like to acknowledge the invaluable help of: the PC members for ensuring the quality of the scientific program, the Tahar Moulay University of Saida and the GeCoDe Laboratory, for hosting the conference and providing all the needed support, the track chairs, for managing the reviewing process, and Dr. Mickael Baron

(ISAE-ENSMA, Poitiers, France) and Dr. Mahieddine Djoudi (SIC/XLIM, University of Poitiers, France) for Webmaster efforts. Last but not least, we thank the EasyChair team for making available their conference management system to CIIA. Finally, we thank the authors who submitted papers to the conference.

April 2015

Abdelmalek Amine
Erich J. Neuhold
Ladjel Bellatreche
Zakaria Elberrichi
Robert Wrembel

Organization

CIIA 2015 was organized by the GeCoDe Laboratory and Tahar Moulay University of Saida (Algeria) in cooperation with the International Federation for Information Processing (IFIP).

Conference Committees

Honorary Chair

Prof. Fethallah Tebboune	Rector of the Tahar Moulay University of Saida, Algeria

General Chairs

Abdelmalek Amine	Tahar Moulay University of Saida, Algeria
Erich Neuhold	University of Vienna, Austria

PC Chairs

Ladjel Bellatreche	LIAS/ISAE-ENSMA, France
Zakaria Elberrichi	Sidi Bel Abbès University, Algeria
Robert Wrembel	Poznan University of Technology, Poland

Steering Committee

Abdelmalek Amine	Tahar Moulay University of Saida, Algeria
Otmane Ait Mohamed	Concordia University, Canada
Ladjel Bellatreche	ISAE-ENSMA, France
Mahieddine Djoudi	SIC/XLIM, University of Poitiers, France
Carlos Ordonez	University of Houston, USA

Track Chairs

Nadjib Badache	USTHB, CERIST, Algeria
Kamel Barkaoui	CNAM, France
Sadok Ben Yahia	FST Tunis, Tunisia
Jorge Bernardino	ISEC-Polytechnic Institute of Coimbra, Portugal
Alfredo Cuzzocrea	ICAR-CNR and University of Calabria, Italy
Abdelwahab Hamou-Lhadj	Concordia University, Montreal, Canada
Nadjet Kamel	University of Setif, Algeria
Selma Khouri	ESI, Algeria

Program Committee

Track 1: Computational Intelligence

Wiem Abdelbaki	University of Nizwa, Oman
Mustapha Kamel Abdi	University of Oran, Algeria
Reda Adjoudj	University of Sidi Bel Abbès, Algeria
Abbes Amira	Qatar University, Quatar and University of the West of Scotland, UK
Saliha Aouat	USTHB, Algeria
Sabeur Aridhi	University of Trento, Italy
Sarah Ayouni	ESIG Kairouan, Tunisia
Latifa Baba-Hamed	University of Oran, Algeria
Ghalem Belalem	University of Oran 1, Algeria
Sid Ahmed Ben Abderrahmane	University Paris 8, France
Mohamed Ben Mohamed	University of Constantine 2, Algeria
Mohamed Chaouki Babahenini	University of Biskra, Algeria
Nadia Baha Touzene	USTHB, Algeria
Mohamed Batouche	University of Constantine, Algeria
Leila Ben Othman	IPEI El Manar, Tunisia
Ismail Biskri	Université du Québec à Trois-Rivieres, Canada
Lydia Boudjeloud	University of Lorraine, France
Aoued Boukelif	University of Sidi Bel Abbès, Algeria
Belattar Brahim	University of Batna, Algeria
Imen Brahmi	FST, Tunisia
Hanen Brahmi	ESIG Kairouan, Tunisia
Laurence Capus	University Laval, Canada
Allaoua Chaoui	University of Constantine 2, Algeria
Salim Chikhi	University of Constantine 2, Algeria
Gayo Diallo	University of Bordeaux, France
Yassine Djouadi	USTHB University, Algeria
Narjes Doggaz	FST, Tunisia
Bourennane El-Bay	University of Bourgogne, France
Samir Elloumi	Qatar University, Qatar
Kamel Mohamed Faraoun	University of Sidi Bel Abbès, Algeria
Cherif Fodil	University of Biskra, Algeria
Ahmed Guessoum	USTHB, Algeria
Zahia Guessoum	University of Paris 6, France
Allel Hadj-Ali	LIAS/ISAE-ENSMA, France
Tarek Hamrouni	ISAM Manouba, Tunisia
Chihab Hanachi	IRIT, France
Salima Hassas	University of Lyon 1, France
Tutut Herawan	University of Malaya, Malaysia
Ali Mohamed Jaoua	Qatar University, Qatar
Warith Eddine Jeddi	Computer Science Institute of Kasserine, Tunisia

Nidhal Jelassi	FST, Tunisia
Marouen Kachroudi	Computer Science Institute of Kef, Tunisia
Samir Kechid	USTHB, Algeria
Hamamache Kheddouci	University of Lyon 1, France
Slimane Larabi	USTHB, Algeria
Phayung Meesad	KMUTNB, Thailand
Mohamed El Bachir Menai	King Saud University, Saudi Arabia
Hayet Merouani	University of Annaba, Algeria
Souhal Meshoul	University of Constantine, Algeria
Takao Miura	Hosei University, Japan
Abdelouahab Moussaoui	University of Setif, Algeria
Kazumi Nakamatsu	University of Hyogo, Japan
Binod Kumar Prasad	Maharashtra Academy of Engineering, India
Mohamed Quafafou	Aix-Marseille University, France
Abdellatif Rahmoun	University of Sidi Bel Abbès, Algeria
Sivaram Rajeyyagari	JNTUK, India
Olivier Raynaud	University of Blaise Pascal, France
Zaidi Sahnoun	University of Constantine, Algeria
Abdel-Badeeh Salem	Ain Shams University, Egypt
Ahmed Samet	Compiegne University, France
Minyar Sassi Hidri	Tunis El Manar University, Tunisia
Aymen Sellaouti	INSAT Tunis, Tunisia
Mohamed Senouci	University of Oran 1, Algeria
Hamid Seridi	University of Guelma, Algeria
Noria Taghezout	University of Oran, Algeria
Chiraz Trabelsi	ISAM Manouba
Taoufik Yeferny	ISLAIB of Béja, Tunisia
Bing Zhou	Sam Houston State University, USA

Track 2: Security and Network Technologies

Maurizio Atzori	University of Cagliari, Italy
Mohamed Aissani	EMP, Algeria
Makhlouf Aliouat	University of Setif 1, Algeria
Abderrahmane Amrouche	USTHB, Algeria
Nadjib Badache	USTHB, CERIST, Algeria
Mouloud Bagaa	CERIST, Algeria
Ghalem Belalem	University of Oran, Algeria
Yacine Belhoul	CERIST, Algeria
Mohamed Benmohamed	Constantine 2 University, Algeria
Chafika Benzaid	USTHB, Algeria
Abdelmadjid Bouabdallah	UTC, France
Yacine Challal	ESI, Algeria
Alfredo Cuzzocrea	ICAR-CNR and University of Calabria, Italy
Abdelouahid Derhab	CoEIA, Saudia Arabia

Djamel Djenouri	CERIST, Algeria
Josep Domingo-Ferrer	Universitat Rovira i Virgili, Spain
Nacira Ghoualmi-Zine	University of Annaba, Algeria
Mohamed Guerroumi	USTHB, Algeria
Abdelkrim Hamza	USTHB, Algeria
Michal Kalewski	Poznan University of Technology, Poland
Anna Kobusinska	Poznan University of Technology, Poland
Noureddine Lasla	CERIST, Algeria
Sekhri Larbi	University of Oran, Algeria
Giovanni Livraga	Università degli Studi di Milano, Italy
Sadegh Nobari	Skoltech Faculty, Russia
Mustapha Reda Senouci	EMP, Algeria
Anna Squicciarini	Pennsylvania State University, USA
Djamel Tandjaoui	CERIST, Algeria
Traian Marius Truta	Northern Kentucky University, USA
Dariusz Wawrzyniak	Poznan University of Technology, Poland
Ali Yachir	EMP, Algeria
Said Yahiaoui	CERIST, Algeria
Youcef Zafoune	USTHB, Algeria

Track 3: Information Technologies

Samir Aknine	University of Lyon 1, France
Ana Almeida	Polytechnic of Porto, Portugal
Karima Amrouche	Ecole Supérieure d'Informatique, Algeria
Witold Andrzejewski	Poznan University of Technology, Poland
Baghdad Atmani	University of Oran 1, Algeria
Bartosz Bebel	Poznan University of Technology, Poland
Orlando Belo	University of Minho, Portugal
Djamal Benslimane	Lyon 1 University, France
Sidi Mohamed Benslimane	University of Sidi Bel Abbès, Algeria
Karim Bouamrane	University of Oran 1, Algeria
Kamel Boukhalfa	USTHB University, Algeria
Nabila Bousbia	Ecole Supérieure d'Informatique, Algeria
Omar Boussaid	University of Lyon 2, France
Zouhaier Brahmia	Faculty of Economics and Management, Tunisia
Rachid Chalal	Ecole Supérieure d'Informatique, Algeria
Abderrahim El-Qadi	University of Moulay Ismail, Morocco
Marcin Gorawski	Silesian University, Poland
Reda Mohamed Hamou	Tahar Moulay University of Saida, Algeria
Saad Harous	UAE University, United Arab Emirates
Walaid Khaled Hidouci	Ecole Supérieure d'Informatique, Algeria
Abdessamad Imine	Loria Nancy, France
Stéphane Jean	LIAS/ISAE-ENSMA, France

Benatchba Karima	Ecole Supérieure d'Informatique, Algeria
Adel Kermi	Ecole Supérieure d'Informatique, Algeria
Ahmed Lehireche	University of Sidi Bel Abbès, Algeria
Moussa Lo	University of Gaston Berger, Senegal
Mimoun Malki	University of Sidi Bel Abbès, Algeria
Elio Masciari	Consiglio Nazionale delle Ricerche, Italy
Elsa Negre	University of Paris - Dauphine, France
Oscar Romero	Universitat Politecnica de Catalunya, Spain
Paolo Rosso	Universitat Politecnica de Valencia, Spain
Hala Skaf-Molli	University of Nantes, France
Rafael Tolosana	Universidad de Zaragoza, Spain
Satya Valluri	Oracle, USA
Leandro-Krug Wives	UFRGS, Brazil
Marek Wojciechowski	Poznan University, Poland
Leila Zemmouchi-Ghomari	USTHB University, Algeria

Track 4: Software Engineering

Mohamed Ahmed-Nacer	USTHB, Algiers, Algeria
Yamine Ait Ameur	ENSEEIHT, Toulouse, France
Hassane Alla	GIPSA, UJF Grenoble, France
Maria-Virginia Aponte	CEDRIC, CNAM, France
Mohamed Faouzi Atig	Uppsala University, Sweden
Abdelkrim Amirat	University of Souk Ahras, Algeria
Faiza Belala	University of Constantine 2, Algeria
Kamel Barkaoui	CNAM, France
Belgacem Ben Hedia	LIST-CEA, Saclay, France
Saddek Bensalem	VERIMAG, UJF Grenoble, France
Frederic Boniol	ONERA, Toulouse, France
Thouraya Bouabana Tebibel	ESI, Algiers, Algeria
Ahmed Bouajjani	LIAFA, University Paris 7, France
Hanifa Boucheneb	Polytechnique Montreal, Canada
Nacer Boudjlida	LORIA, University of Lorraine, France
Zizette Boufaida	LIRE, University Constantine 2, Algeria
Mohand Cherif Boukala	USTHB, Algiers, Algeria
Samia Bouzefrane	CEDRIC, CNAM, France
Manfred Broy	TU München, Germany
Christine Choppy	LIPN, University of Paris 13, France
Annie Choquet-Geniet	ISAE-ENSMA, France
Karim Djouani	F'SATI/TUT, Pretoria, South Africa
Amal El Fallah Seghrouchni	LIP6, UPMC, Paris, France
Mohamed Erradi	LAGI, ENSIAS Rabat, Morocco

Alessandro Fantechi	University of Florence, Italy
Mohamed Mohsen Gammoudi	University of Manouba, Tunisia
Faiez Gargouri	MIRACL, ISIM, Sfax, Tunisia
Stefan Haar	LSV- CNRS and ENS Cachan, France
Mohand Said Hacid	LIRIS, University of Lyon 1, France
Henda Hajjami Ben Ghezala	ENSI, Tunisia
Abdelwahab Hamou-Lhadj	Concordia University, Montreal, Canada
Rolf Hennicker	Ludwig-Maximilians-Universität München, Germany
Ali Mohamed Jaoua	Qatar University Qatar
Mohamed Jmaiel	ReDCAD, ENIS, Sfax, Tunisia
Okba Kazar	University of Biskra, Algeria
Anna-Lena Lamprecht	SSE, University of Potsdam, Germany
Zhiwu Li	Xidian University, China
Mourad Maouche	Philadelphia University Amman, Jordan
Tiziana Margaria	University of Limerick and Lero, Ireland
Mohamed Mezghiche	University of Boumerdes, Algeria
Ali Mili	NJIT, Newark, USA
Bruno Monsuez	UIIS, ENSTA ParisTech, France
Mohamed Mosbah	LaBRI, Bordeaux INP, France
Hassan Mountassir	LIFC, University of Franche-Comté, France
Mourad Chabane Oussalah	LINA, University of Nantes, France
Ahmed Rezine	Linköpings Universitet, Sweden
Riadh Robbana	LIP2, INSAT, Tunis, Tunisia
Samir Tata	Telecom SudParis, France
Farouk Toumani	LIMOS, University of Clermont-Ferrand, France
Nadia Zeghib	LIRE, University Constantine 2, Algeria

Local Organizing Committee

Chair: Reda Mohamed Hamou
Webmaster: Mahieddine Djoudi
Mohamed Derkaoui
Mahmoud Fahsi
Toufik Guendouzi
Abdelkader Khobzaoui
Ahmed Chaouki Lokbani
Kheireddine Mekkaoui
Abdelkader Mostefai
Mohamed Rahmani

Sponsoring Institutions

CIIA 2015 received the support of several sponsors, among them Tahar Moulay University of Saida, Algerie, GeCoDe Laboratory of Tahar Moulay University, Saida, IFIP, ISAE-ENSMA, LIAS Laboratory (Poitiers), XLIM/SIC (Poitiers), ARPT, DG-RSDT. Many thanks for their support.

Invited Talks

Interoperability: Models and Semantics - A Reoccurring Problem

Erich J. Neuhold

University of Vienna, Austria
erich.neuhold@univie.ac.at
http://cs.univie.ac.at/Erich.Neuhold

Abstract. Interoperability is a qualitative property of computing infrastructures that denotes the ability of the sending and receiving systems to exchange and properly interpret information objects across system boundaries.

Since this property is not given by default, the interoperability problem involves the representation of meaning and has been an active research topic for approximately four decades. Database models used schemas to express semantics and implicitly aimed at achieving interoperability by providing programming independence of data storage and access.

After a number of intermediate steps such as Hypertext and XML document models, the notions of semantics and interoperability became what they have been over the last ten years in the context of the World Wide Web and more recently the concept of Open Linked Data.

The talk will investigate the (reoccurring) problem of interoperability as it can be found in the massive data collections around the Big Data and Open Linked Data concepts. We investigate semantics and interoperability research from the point of view of information systems. It should give an overview of existing old and new interoperability techniques and point out future research directions, especially for concepts found in Open Linked Data, the Semantic WEB and Big Data.

Brain-Computer-Brain Interfaces for Sensing and Subsequent Treatment

Mohamad Sawan, Professor and Canada Research Chair

Polystim Neurotechnologies Laboratory, Polytechnique Montreal
mohamad.sawan@polymtl.ca

Abstract. Implantable Brain-Computer-Brain Interfaces (BCIs) for diagnostic and recovery of neural vital functions are promising alternative to study neural activities underlying cognitive functions and pathologies. This Keynote address covers the architecture of typical BCI intended for wireless neurorecording and neurostimulation. Massively parallel multichannel spike recording through large arrays of microelectrodes will be introduced. Attention will be paid to low-power mixed-signal circuit design optimization. Advanced signal processing implementation such as adaptive thresholding, spike detection, data compression, and transmission will be described. Also, the talk includes Lab-on-chip technologies intended to build biosensors, and wireless data links and harvesting power to implants. Tests and validation of devices : electrical, mechanical, package, heat, reliability will be summarized. Case studies will be covered and include research activities dedicated to vision recovery through implant used to apply direct electrical microstimulation, to present the environment as phosphenes in the visual field of the blind. And we will summarize latest activities on locating epileptic seizures using multi-modal fNIRS/EEG processing, and will show the onset detecting seizure and techniques to stop it, using bioelectronic implant.

Collaborative and Social Web Search

Lynda Tamine

Université Paul Sabatier, Institut de Recherche en Informatique de Toulouse
Toulouse, France
tamine@irit.fr

Abstract. Web search increasingly reflects problems grounded in the real-life world that requires the assistance of social resources. Social web search refers broadly to 1) the process of searching information over user-generated content (UGC) or 2) searching online with the help of users (such as friends, colleagues or experts) using large-scale social networking services. Examples of such services include Facebook, Twitter and MySpace and are considered as complementary to web search engines. Collaborative search is a kind of social search where small-scale groups of users are all together engaged in solving a shared information need. Collaborative and social search allow the gathering of users' complementary knowledge and skills that lead to the emergence of collective intelligence.

The aim of this talk is to 1) outline the paradigm of social search, 2) investigate the research issues that it gives rise to and then 3) point out the opportunities it brings to nowadays society.

I will look back over the past recent years highlighting some of the major changes in social-centred approaches of information search and related main research findings. I will also give an overview and share some experiences we gained through our previous research investigations in the area of collaborative and social search.

Contents

Computational Intelligence: BioInformatics

Information Technology: Text and Speech Processing

Information Technology: Requirement Engineering

Information Technology: OLAP and Web Services

Information Technology: Recommender Systems and Web Services

Information Technology: Ontologies

Security and Network Technologies: Security

Security and Network Technologies: Wireless Sensor Networks

Security and Network Technologies: Energy and Synchronisation

Security and Network Technologies: Potpourri

Software Engineering: Modeling and Meta Modeling

Software Engineering: Checking and Verification

Computational Intelligence: Meta-heuristics

Binary Bat Algorithm:
On The Efficiency of Mapping Functions When Handling Binary Problems Using Continuous-variable-based Metaheuristics

Zakaria Abd El Moiz Dahi(✉), Chaker Mezioud, and Amer Draa

Modeling and Implementation of Complex Systems laboratory
Dept. of New Technologies of Information and Communication
Constantine 2 university
Constantine City, Algeria
{zakaria.dahi,chaker.mezioud}@univ-constantine2.dz,
{draa_amer@yahoo.fr}

Abstract. Global optimisation plays a critical role in today's scientific and industrial fields. Optimisation problems are either continuous or combinatorial depending on the nature of the parameters to optimise. In the class of combinatorial problems, we find a sub-category which is the binary optimisation problems. Due to the complex nature of optimisation problems, exhaustive search-based methods are no longer a good choice. So, metaheuristics are more and more being opted in order to solve such problems. Some of them were designed originally to handle binary problems, whereas others need an adaptation to acquire this capacity. One of the principal adaptation schema is the use of a mapping function to decode real-valued solutions into binary-valued ones. The Antenna Positioning Problem (APP) is an NP-hard binary optimisation problem in cellular phone networks (2G, EDGE, GPRS, 3G, 3G$^+$, LTE, 4G). In this paper, the efficiency of the principal mapping functions existing in the literature is investigated through the proposition of five binary variants of one of the most recent metaheuristic called the Bat Algorithm (BA). The proposed binary variants are evaluated on the APP, and have been tested on a set of well-known benchmarks and given promising results.

1 Introduction

Combinatorial problems are problems whose parameters belong to a finit set of integers ($x_i \epsilon \mathbb{N}$). The latter includes a more specific type called *binary optimisation problems* : problems whose parameters can take values from a bi-valued search space called *genotype space* ($x_i \epsilon \{ 1 , 0 \}$).

The design of cellular phone networks (2G, EDGE, GPRS, 3G, 3G$^+$, LTE, 4G) is one of the most critical tasks during the network implantation. Any design process that can not deal with this phase may alter the service quality

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 3–14, 2015.
DOI: 10.1007/978-3-319-19578-0_1

of the network itself. The *Antenna Positioning Problem* (*APP*) is one of the most challenging optimisation issues in the design phase of cellular networks. The APP is formulated as a binary optimisation problem and was proven to be NP-hard.

Metaheuristics are efficient tools to use when tackling such optimisation problems. Regardless to the source of their inspiration, metaheuristics can be divided into algorithms who are originally designed to tackle continuous problems, and those who are designed to tackle combinatorial ones.

The *Bat Algorithm* (*BA*), is one of the recently proposed metaheuristics [14]. It was inspired by the natural phenomenon of echolocation used by bats. The BA was originally designed to tackle optimisation problems within continuous search space and it has shown encouraging performances.

Generally, when adapting a *continuous-variable-based metaheuristic* (i.e. *a metaheuristic that was designed originally to operate on variables within continuous search space*) to tackle binary problems, many schemas of adaptation exist. One of the most opted one, is the use of a mapping function to map the real-valued solutions into binary ones. Several sub-schemas of mapping exist as well in this last one : *one-to-one*, *many-to-one*, *one-to-many*.

Many questions still surround this schema of adaptation, such as the fact that the efficiency of these mapping functions is still fuzzy and unexplored. In addition, no clear statement exist on whether using *binary-variable-based metaheuristics* (i.e. *metaheuristics that were designed originally to operate on variables within binary-valued search space*) or continuous ones to tackle binary optimisation problems. Does the efficiency of these mapping functions depends on the algorithm used or the problem solved. Finally, no affined study shows if it is worth using this mapping functions and ultimately which kind of metaheuristics is more efficient when solving binary problems.

Through analysing the literature, the five principal mapping functions existing are used to propose new binary variants of the *Bat Algorithm*. The mapping functions used in this work were selected to illustrate different schemas of mapping. These functions are : The *Nearest Integer method* (*NI*), the *normalisation technique*, the *Angle Modulation method* (*AM*), the *Great Value Priority method* (*GVP*) and finally, the *Sigmoid Function* (*SF*). The proposed binary variants of the BA were tested on the APP using well-known benchmark instances, with different sizes and complexity as well. In addition, the last ones were compared to one of the most used binary-variable based algorithm : the *Genetic Algorithm*.

The remainder of this paper is structured as follows. In Section 2, we introduce basic concepts related to the antenna positioning problem, the bat algorithm, and the mapping functions used in this work. In Section 3, we introduce the proposed binary variants of the bat algorithm. Section 4 is dedicated to experimental results, their interpretation and disscussion. Finally, we present the conclusion of our work in Section 5.

2 Basic Concepts

In this Section, we introduce basic concepts related to the antenna positioning problem, the bat algorithm, and the used mapping functions.

2.1 Antenna Positioning Problem

In this section, we present a widely used formulation of the APP, that was given by Guidec et. al. [3]. This modeling of the antenna positioning problem consider two objectives : maximizing the covered area while minimizing the number of base station used.

The antenna positioning problem recalls NP-hard problems in graph theory such as the Minimum Dominating Set (MDS), the Maximum Independent Set (MIS), or the Unicast Set Covering Problem (USCP), see Sub-figures (a), (b), (c) of Figure 2.

Let L be the set of all potentially covered areas and M the set of all potential locations of base stations. Suppose G is a graph where E is the set of edges in the graph verifying that each transmitter is linked to the area that it covers. One seeks for the minimum subset of transmitters that covers the maximum surface of a given area. In other words, the objective is to find a subset M' such that $|M'|$ is minimised and $\mid Neighbours(M', E) \mid$ is maximised [4]; where $Neighbours$ represents the set of the covered area, and M' represents the set of transmitters used to cover this area.

$$\mid Neighbours(M', E)\mid \; = \; \{\, u \,\epsilon\, L \,\mid \exists\, v \,\epsilon\, M' \,, (u, v) \,\epsilon\, E \,\} \tag{1}$$

$$M' \; = \; \{\, t \,\epsilon\, B \,\mid x_t \; = \; 1 \,\} \tag{2}$$

A Base Tranceiver Station (BTS) is a radio transmitting device with a specific type of coverage (See Figure 1). In this work, we used three types of coverage introduced in [1]. A cell is a part of a geographical area that is covered by a base station.

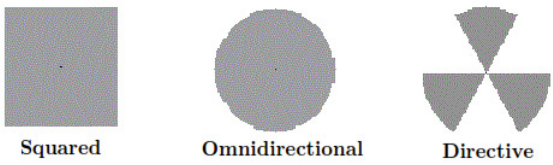


Fig. 1. Antenna Coverage Models

The working area is discretized in a rectangular grid with Dim_x and Dim_y dimensions. Having $Sites \; = \; \{site_1 \,,\, site_2 \,,\, site_3 \,, \ldots \,,\, site_N \,\}$ is the set of potentially preselected sites, where the antennas can be placed. Each potential

site location is identified by Cartesian coordinates $\{ site_1 = (x_1 , y_1) , site_2 = (x_2 , y_2) , site_3 = (x_3 , y_3) , \ldots , site_N = (x_N , y_N) \}$, see Sub-figures (d), (e), (f) of Figure 2.

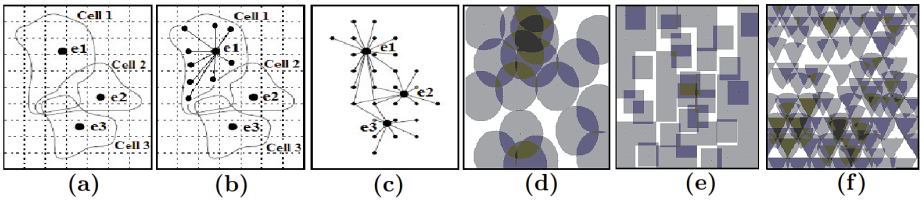


Fig. 2. Representation of The Discretized Area

A potential solution of the APP can be a binary vector described as follows. Each vector $\overrightarrow{X}$ represents a potential configuration of the mobile network. The number of elements of each vector represents the number of potential condidate sites. The rank of each dimsenion represents the rank of the corresponding base station $i = 1 , 2 , \ldots , Dimension_{\overrightarrow{x}}$. Each dimension of the vector is strictly binary valued : $x_i \epsilon \overrightarrow{X} / x_i = 1 \vee x_i = 0$. If $x_i = 1$, the i^{th} base station is selected, otherwise it is discarded. The objective function to optimize is defined by the formula 3.

$$Maximize : \quad f(x) = \frac{Cover\ ratio^{\alpha}}{Number\ of\ used\ base\ station} \tag{3}$$

with :

$$Number\ of\ used\ base\ station = \sum_{i=1}^{Dimension_{\overrightarrow{x}}} x_i \ , \tag{4}$$

with :

$$cover\ ratio = \left(\frac{Covered\ area}{Total\ area} \right) * 100 \tag{5}$$

And :

$$Covered\ area = \sum_{i=1}^{Dim_x} \sum_{j=1}^{Dim_y} cover(i,j) \ , \tag{6}$$

And :

$$Total\ area = Dim_x * Dim_y . \tag{7}$$

It is worth to mention that other mathematical models of the antenna positioning problem exist like the one proposed in [12]. Generally, these models differ by their mathematical formulations or modeling.

2.2 Bat Algorithm

The Bat Algorithm has been recently proposed. It is a swarm-based metaheuristic [14]. This algorithm is inspired by the natural echolocation behaviour of bats. Microbats use a type of sonar, called echolocation, to detect their preys, avoid obstacles, and locate their roosting crevices in the dark. These bats emit a very loud sound pulse and listen for the echo that bounces back from the surrounding objects. Their pulses vary in properties and can be correlated with their hunting strategies, depending on the species. The bats then adjust the pulse and rate of the sound as they get closer to the obstacles or the prey. This phenomenon has been translated into the newly proposed bat algorithm. The pseudo-code of Algorithm 1 describes the general framework of the bat algorithm.

Algorithm 1. The Bat Algorithm

1. Objective function $f(x), x = (x_1, ..., x_d)^T$
2. Initialize the bat population X_i (i = 1, 2, ..., n) and V_i
3. Define the pulse rate r_i and the loudness A_i
4. **Input :** Initial bat population
5. **while** ($t <$ Max number of iterations) **do**
6. Generate new solutions by adjusting frequency, and updating velocities and locations/solutions (Equations 8 to 10)
7. **if** ($rand > r_i$) **then**
8. Select a solution among the best solutions
9. Generate a local solution around the selected best solution
10. **end if**
11. Generate a new solution by flying randomly
12. **if** ($rand < A_i \,\&\, f(X_i) < f(X_*)$) **then**
13. Accept the new solutions
14. Increase r_i and reduce A_i
15. **end if**
16. Rank the bats and find the current best X_*
17. **end while**
18. **Output :** Best bat found (i.e. best solution)

Equations 8, 9 and 10 define how the position X_i and velocity V_i in a d-dimensional search space are updated. The new solution X_i^t and velocity V_i^t at a time step t are given by :

$$f_i = f_{min} + (f_{max} - f_{min})\beta \tag{8}$$

$$V_i^t = V_i^{t-1} + (X_i^t - X_*)f_i \tag{9}$$

$$X_i^t = X_i^{t-1} + V_i^t \tag{10}$$

Initially, each bat is randomly assigned a frequency which is drawn randomly from $[f_{min}, f_{max}]$. $\beta \epsilon [0, 1]$ is a random vector drawn from a uniform distribution. X_* is he current global best location (solution) which is located after comparing all the solutions.

When a local search is performed a solution is selected among the current best solutions. A new solution for each bat is generated using a random walk as described in Equation 11; Where ϵ is a random number from $[-1, 1]$, and $A^t =< A_i^t >$ is the average loudness of all bats at this time step.

$$X_{new} = X_{old} + \epsilon A^t \tag{11}$$

Likewise, the loudness A_i and the rate ri of pulse are updated once the new solution is accepted. This is done using Formulas 12 and 13, where α and γ are constants. Initially, each bat is randomly assigned a loudness and a rate drawn respectively from the intervals $[A_{min}, A_{max}]$ and $[r_{min}, r_{max}]$.

$$A_i^{t+1} = \alpha A_i t \tag{12}$$

$$r_i^{t+1} = r_i^0 [1 - exp(-\gamma t)] \tag{13}$$

2.3 From Phenotype to Genotype Space

Several approaches exist for adapting a continuous-variable-based metaheuristic to work also in binary search space. The first schema of adaptation consists in replacing arithmetic operators of the metaheuristic by logical ones to operate directly on the binary solutions. The second aims to find the corresponding operators of the algorithms in the geometric space (*Hamming space*). Finally, the third approach consists in conserving the original operators, architecture and solution representation of the algorithm, and adding a complementary module that maps real-valued solutions into binary ones. The techniques used in this third category are generally known also as *mapping functions*. Several schemas of mapping exist. The pincipal ones are : *one-to-one*, *many-to-one*, and *one-to-many*.

The mapping functions used in this work have been chosen to illustrate several mapping schemas and several mathematical properties. In the following, the mapping functions used in this study are introduced.

Nearest Integer (NI): This technique consists of assigning a real number to the nearest integer by rounding or truncating it up or down [2,6].

Normalisation: This approach was proposed in [9,10]. It consists of the normalisation of the solution by linearaly scaling it using the Formula 14. Then the condition 15 is applied to get the corresponding binary solution.

$$x_{ij}^{'} = \frac{(\, x_{ij} + x_i^{min}\,)}{(\,|\, x_i^{min} \,| + x_i^{max}\,)} \tag{14}$$

$$x_{ij} = \begin{cases} 1\,, & If\ x_{ij}^{'} \geqslant 0.5 \\ 0\,, & Otherwise \end{cases} \tag{15}$$

Assuming that $i = 1 \cdots N$ and $j = 1 \cdots D$. Where : N is the population size, D is the size of the solution vector. x_i^{min} and x_i^{max} are respectively the minimum and the maximum values in the i^{th} vector at the iteration t.

Angle Modulation (AM): The idea is to use a trigonometric function to map real-valued solutions into binary ones. The generator function is used for signal processing in telecommunications and defined as follows [11,13].

$$g(x_{ij}) = sin\,(\,2\Pi\,(\,x_{ij} - a)\,*\,b\,*\,cos\,(\,2\Pi\,(\,x_{ij} - a)\,*\,c\,)\,)\,+\,d \tag{16}$$

Where $i = 1 \cdots N$ and $j = 1 \cdots D$. N is the population size, D is the size of the solution vector. $g(x)$ is the generator function, and x_{ij} is a single element from a potential solution vector. Instead of optimizing a D-dimensional binary string solution, the search space is reduced to a 4-dimensional search space. Each vector of solution $\overrightarrow{G}$ represents potential values of the coefficients (a, b, c, d) in the generator function. At each iteration, every solution vector is applied to a sample vector $\overrightarrow{X}$ with the original D-dimensions of the problem. The sample vector is drawn from a uniform distribution and has equaly spaced intervals between each dimension and another. Finally, one has to apply the following formula on the resulting vector :

$$x_{ij} = \begin{cases} 1\,, & If\ g(x_{ij}) \geqslant 0 \\ 0\,, & Otherwise \end{cases} \tag{17}$$

Sigmoid Function (SF): In this technique, each real valued dimension of the solution vector is mapped into a strictly binary valued one [7,8]. The probability of each dimension to flip to one state or another is computed according to the real value of the dimension itself by applying Formula 18.

$$x_{ij} = \begin{cases} 1\,, & If\ Rand[0,1] \leq \frac{1}{1+e^{x_{ij}}} \\ 0\,, & Otherwise \end{cases} \tag{18}$$

Where $i = 1 \cdots N$ and $j = 1 \cdots D$. N is the population size, D is the size of the solution vector, and $Rand[0,1]$ is a randomly generated positive number, drawn from a uniform distribution in the interval $[\,0\,,1\,]$.

Great Value Priority (GVP): Recently, authors in [5] have introduced this technique. Starting from a given real valued solution vector $\overrightarrow{X}$, a permutation vector $\overrightarrow{P}$ is created. The first element of the permutation vector p_1 will contain the position of the largest element in the original vector, the second element of the permutation vector p_2 will receive the position of the second largest element of the real valued vector, and so on. The procedure will be repeated until all the elements of the original vector are browsed. Finally, having the permutation vector $\overrightarrow{P}$, the following formula will be applied to recover back a binary valued vector.

$$x_{ij} = \begin{cases} 1\,, & If\ p_j\ >\ p_{j+1} \\ 0\,, & Otherwise \end{cases} \tag{19}$$

3 The proposed Binary Bat Algorithm

The inclusion of the discretisation step using one of the mapping functions after line code 10 in pseudo-code of Algorithm 1 results in giving birth to new variants of the bat algorithm. The first variant, using the nearest ineteger method as a discretising technique is called NI-BBA (for Nearest Integer based Binary Bat Algorithm). The second variant is called N-BBA (for Normalisation based Binary Bat Algorithm) is based on the normalisation method. The third variant uses the sigmoid function and is called SF-BBA (for Sigmoid Function based Binary Bat Algorithm). The fourth variant is the AM-BBA (for Angle Modulation based Binary Bat Algorithm) is based on the angle modulation method. Finally, GVP-BBA (for Great Value Priority based Binary Bat Algorithm) uses the great value priority technique.

4 Experimental Results and Discussion

The experiments where carried using an Intel I3 core with 2 GB Ram and a Windows 7 OS. The implementation was done using Matlab 7.12.0 (R2011a).

Two scenarios were randomly generated. Both are representing a working area of 20.25 Km^2. The first instance contains 549 available locations, whereas the second instance contains 749 available locations. Other instances of 149 and 349 preselected positions, are used here. They were provided by the university of Malaga, Spain. We used three types of coverage : squared, omnidirectional and directive [1]. It is worth to mention that directive antennas cover one sixth of the area of omnidirectionnal antennas having the same radius. Table 1 shows all the features of the used instances.

The proposed binary variants of the bat algorithm were also compared to one of the most used binary-variable-based metaheuristic and whose the efficiency is well established : the canonical Genetic Algorithm (GA). The GA uses a wheel selection and a two-point crossover with a probability equal to 0.7 and a bit-flip

Table 1. Instances : Size and Coverage

Instance Type	Grid Dimension	Instance	Coverage	Radius
Synthetic	287 x 287	149	Omnidirectional	22
			Squared	20
			Directive	22
		349	Omnidirectional	22
			Squared	20
			Directive	22
Random	300 x 300	549	Omnidirectional	26
			Squared	24
			Directive	26
		749	Omnidirectional	26
			Squared	24
			Directive	26

mutation with a probability of 0.05. The percentage of chromosomes used to create the matting pool is 50 %. The parameters of the bat algorithm used in this experiment are shown in Table 2.

Table 2. Bat Algorithm Parameters

Prameter	Value
f_{max}	10
f_{min}	-10
A_{max}	2
A_{min}	1
r_{max}	1
r_{min}	0
α	0.9
γ	0.9

The experiments were performed till reaching 20.000 evaluations, and each one is repeated for 20 runs. Several results are reported such as : the *best* and the *worst* fitness, and also the *mean* and *standard deviation* of fitness value over 20 runs.

Tables 3, 4, 5 and 6 show the results obtained when evaluating the five binary variants of the bat algorithm using the instances 149, 349, 549, 749 and this for each type of coverage : squared, omnidirectional and directive.

Based on the results shown in Tables 3, 4, 5 and 6 many observations can be made. The performances of the variants for small instances (149, 349) are close, but one can note that as the instance size increases (549, 749), the difference in the efficiency of the variants is more obvious. So, for some variants like the AM-BBA, N-BBA, GVP-BBA their efficiency depends highly on the size of the problem treated. Whereas for other variants such as the NI-BBA and the SF-BBA the efficiency is maintained even if the size of problem increases.

In general, NI-BBA and SF-BBA variants has shown better results than the other variants when solving the APP, especially the NI-BBA. The scalability of both variants is similar since both succeeded to solve the different sizes of

Table 3. Results of the Bat Algorithm Variants For Instance 149

Instance	Coverage	Algorithm	Best	Worst	Mean	Std
149	Squared	NI-BBA	120.582	120.582	**120.582**	1.458E-14
		N-BBA	106.361	95.006	100.501	2.70580121
		AM-BBA	111.120	97.564	104.250	4.63919122
		SF-BBA	103.012	102.112	102.157	0.20130794
		GVP-BBA	113.548	113.400	113.489	0.0745577
		GA	110.495	99.044	104.332	3.60183932
	Circle	NI-BBA	97.701	97.701	97.701	2.916E-14
		N-BBA	94.310	87.932	90.455	1.7646704
		AM-BBA	99.283	85.004	90.940	3.79249425
		SF-BBA	100.366	100.366	**100.366**	2.916E-14
		GVP-BBA	98.747	98.747	98.747	2.916E-14
		GA	97.282	85.472	90.832	2.99547967
	Directive	NI-BBA	41.473	41.473	41.473	1.458E-14
		N-BBA	40.354	37.315	38.905	0.81358109
		AM-BBA	42.560	40.963	41.594	0.44465264
		SF-BBA	42.388	41.859	**42.362**	0.11845141
		GVP-BBA	40.639	40.639	40.639	0
		GA	41.543	36.904	38.924	1.10100813

Table 4. Results of the Bat Algorithm Variants For Instance 349

Instance	Coverage	Algorithm	Best	Worst	Mean	Std
349	Squared	NI-BBA	95.371	95.371	95.371	2.916E-14
		N-BBA	61.664	58.335	59.981	1.03311284
		AM-BBA	188.758	89.671	**135.391**	30.2152069
		SF-BBA	102.880	98.142	102.643	1.05964286
		GVP-BBA	63.551	62.123	63.479	0.319142
		GA	63.643	57.858	61.196	1.60235842
	Circle	NI-BBA	95.081	95.081	**95.081**	2.916E-14
		N-BBA	60.350	56.624	58.753	0.98804881
		AM-BBA	127.577	75.920	88.332	13.1198896
		SF-BBA	90.144	88.803	88.911	0.3445047
		GVP-BBA	61.270	60.903	61.251	0.08199193
		GA	62.199	57.224	59.614	1.48976652
	Directive	NI-BBA	41.464	41.464	**41.464**	7.29E-15
		N-BBA	39.659	37.102	38.221	0.54525135
		AM-BBA	42.283	38.420	39.408	0.91636124
		SF-BBA	39.899	39.899	39.899	7.29E-15
		GVP-BBA	39.723	39.723	39.723	7.29E-15
		GA	39.450	37.199	38.304	0.66880293

the problem. No clear conclusion can be made about how the proposed variants behave when dealing with a specific type of coverage (squared, omnidirectional, directive), or a specific type of data (random, synthetic). One can note also that all the binary variants of the BA were able to outperform the results obtained by the canonical GA for all the instances and for all the sizes of the instances. But one can note also that the difference between the GA and the other variants decreases as the size of the instance decreases.

The conclusion that can be made concerning the impact of the mapping functions on the efficiency of an algorithm, is that the adequate use of a mapping function depends in some cases on the size of the problem engaged and in other cases on the type of the problem. Furthermore, one can note that in reality the bat algorithm do not need complex mapping functions since the basic *rounding function* has shown better results than the other complex mapping functions.

Table 5. Results of the Bat Algorithm Variants For Instance 549

Instance	Coverage	Algorithm	Best	Worst	Mean	Std
549	Squared	NI-BBA	139.973	139.973	139.973	2.916E-14
		N-BBA	40.668	38.939	40.063	0.46911297
		AM-BBA	134.322	96.348	112.442	8.45490438
		SF-BBA	147.445	147.445	**147.445**	2.916E-14
		GVP-BBA	40.529	40.365	40.521	0.03669093
		GA	42.777	38.927	41.070	1.19535659
	Circle	NI-BBA	127.289	126.025	**126.910**	0.59402311
		N-BBA	41.311	38.693	39.740	0.77547919
		AM-BBA	116.780	89.794	102.404	6.32663866
		SF-BBA	117.067	116.067	117.017	0.223514
		GVP-BBA	41.411	41.411	41.411	7.29E-15
		GA	42.640	39.027	40.620	1.06845162
	Directive	NI-BBA	49.046	49.046	**49.046**	1.458E-14
		N-BBA	36.371	34.536	35.468	0.4751343
		AM-BBA	52.156	45.135	48.721	1.83706437
		SF-BBA	51.874	51.808	51.870	0.01464278
		GVP-BBA	35.853	35.814	35.852	0.00873771
		GA	37.913	34.444	35.851	0.89470107

Table 6. Results of the Bat Algorithm Variants For Instance 749

Instance	Coverage	Algorithm	Best	Worst	Mean	Std
749	Squared	NI-BBA	135.888	135.888	**135.888**	2.92E-14
		N-BBA	29.847	28.486	29.175	0.40036354
		AM-BBA	126.827	90.586	109.403	11.4041269
		SF-BBA	130.915	130.915	130.915	0
		GVP-BBA	29.586	29.586	29.586	3.65E-15
		GA	30.788	28.477	29.437	0.52843974
	Circle	NI-BBA	101.870	101.870	101.870	1.46E-14
		N-BBA	29.917	28.275	29.018	0.45477513
		AM-BBA	110.941	87.658	97.607	6.66552368
		SF-BBA	114.077	114.077	**114.077**	0
		GVP-BBA	29.367	29.142	29.356	0.05020659
		GA	30.579	27.929	29.198	0.82131334
	Directive	NI-BBA	50.405	50.024	50.386	0.08519864
		N-BBA	28.674	27.113	27.730	0.4651209
		AM-BBA	50.920	45.767	48.605	1.36744246
		SF-BBA	51.189	49.877	**51.123**	0.29334724
		GVP-BBA	27.578	27.415	27.503	0.02651541
		GA	29.608	26.802	27.785	0.66481778

5 Conclusion

In this paper we conducted a comparative study on the impact of mapping functions on the efficiency of the continuous-variable-based metaheuristic. This was done by proposing five new variants of a recent metaheuristic which is the Bat Algorithm (BA). The proposed binary variants, were tested on an NP-hard optimisation problem in cellular phone networks which is the Antenna Positioning Problem (APP). The results showed that the impact of such mapping functions on the efficiency of an algorithm depends on two factors : the size of the problem, or the complexity of the problem. The best mapping functions found for the bat algorithm are the nearest integer and sigmoid function techniques.

This work illustrates a simple comparative study, and no deep and general conclusion can be made about the efficiency of the mapping functions, the controlling

factor of these last ones or the usefulness of these mapping functions when solving binary problems. So, we seek to conduct a more deep statistical comparative study using several continuous-variable-based metaheuristics and compare them with more powerful binary-variable-based metaheuristics.

References

1. Alba, E., Molina, G., Chicano, J.F.: Optimal placement of antennae using metaheuristics. In: Boyanov, T., Dimova, S., Georgiev, K., Nikolov, G. (eds.) NMA 2006. LNCS, vol. 4310, pp. 214–222. Springer, Heidelberg (2007)
2. Burnwal, S., Deb, S.: Scheduling optimization of flexible manufacturing system using cuckoo search-based approach. The International Journal of Advanced Manufacturing Technology 64(5-8), 951–959 (2013)
3. Calegari, P., Guidec, F., Kuonen, P.: A parallel genetic approach to transceiver placement optimisation. In: Proceedings of the SIPAR Workshop: Parallel and Distributed Systems, pp. 21–24 (1996)
4. Calegari, P., Guidec, F., Kuonen, P., Kobler, D.: Parallel island-based genetic algorithm for radio network design. J. Parallel Distrib. Comput. 47(1), 86–90 (1997)
5. Congying, L., Huanping, Z., Xinfeng, Y.: Particle swarm optimization algorithm for quadratic assignment problem. In: Proceedings of the International Conference on Computer Science and Network Technology (ICCSNT), vol. 3, pp. 1728–1731 (2011)
6. Costa, M., Rocha, A.A.M., Francisco, B.R., Fernandes, M.E.: Heuristic-based firefly algorithm for bound constrained nonlinear binary optimization. Advances in Operations Research 1(215182), 12 (2014)
7. Liu, Q., Lu, W., Xu, W.: Spectrum allocation optimization for cognitive radio networks using binary firefly algorithm. In: Proceedings of the International Conference on Innovative Design and Manufacturing (ICIDM), pp. 257–262 (2014)
8. Palit, S., Sinha, S., Molla, M., Khanra, A.: A cryptanalytic attack on the knapsack cryptosystem using binary firefly algorithm. In: Proceedings of the 2nd International Conference on Computer and Communication Technology (ICCCT), pp. 428–432 (2011)
9. Pampara, G., Engelbrecht, A.: Binary artificial bee colony optimization. In: Proceedings of the IEEE Symposium on Swarm Intelligence (SIS), April 11-15, pp. 1–8 (2011)
10. Pampara, G., Engelbrecht, A., Franken, N.: Binary differential evolution. In: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2006, pp. 1873–1879 (2006)
11. Pampara, G., Franken, N., Engelbrecht, A.: Combining particle swarm optimisation with angle modulation to solve binary problems. In: Proceedings of the Congress on Evolutionary Computation, CEC 2005, pp. 89–96 (2005)
12. Segura, C., Segredo, E., González, Y., León, C.: Multiobjectivisation of the antenna positioning problem. In: International Symposium on Distributed Computing and Artificial Intelligence (DCAI), pp. 319–327 (2011)
13. Swagatam, D., Rohan, M., Rupam, K., Thanos, V.: Multi-user detection in multicarrier CDMA wireless broadband system using a binary adaptive differential evolution algorithm. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation, GECCO 2013, pp. 1245–1252 (2013)
14. Yang, X.S.: A new metaheuristic bat-inspired algorithm. In: Nature Inspired Cooperative Strategies for Optimization (NICSO), vol. 284, pp. 65–74. Springer, Heidelberg (2010)

Relative Timed Model for Coordinated Multi Agent Systems

Said Layadi[1(✉)], Jean-Michel Ilié[2], Ilham Kitouni[1], and Djamel-Eddine Saidouni[1]

[1] MISC Laboratory, University of Abdelhamid Mehri – Constantine 2, 25000, Constantine, Algeria
{layadi,kitouni,saidouni}@misc-umc.org
[2] Universities of Paris UPMC and Paris Descartes 4 place Jussieu, 75005, Paris, France
jean-michel.ilie@upmc.fr

Abstract. The MAS engineering is becoming very important, it is concerned with models, methods and tools. Therefore, verifying the correctness of MAS is the next challenge. We are interested by MAS where each participating agent has its own physical clock of varying frequency, while no global clock is available or desirable. Under such circumstances models must be adapted. In this paper we attempt a novel approach to model the MAS, with a respect of two characteristics, the concurrent aspect and heterogeneity of agents (perceived as a different time rates of agents plan execution). Timed automata with action durations are used; for the circumstance it's extended to deal with relative time rates. Its semantic is abstracted by a novel equivalence relation leading to a region automaton for decidability assessment and proof.

Keywords: Relative time rates · Timed automata with action durations · Region graph · Multi agent systems · Timed transport fleets

1 Introduction

Multi Agent Systems (MAS) are ever-present in computer science applications. This paradigm is used in different domains where reactivity, mobility, dynamicity and adaptation of the system to uncertain or unpredictable factors should be considered. The MAS engineering (i.e. specification, development, management, deployment…) is becoming very important. It is concerned with models, methods and tools. Therefore, verifying the correctness of MAS becomes a fantastic challenge.

We are interested by MAS where each participating agent has its own physical clock of varying frequency, while no global clock is available or desirable. Under such circumstances it's impossible to model system using discrete semantics of time without considering the clock frequencies of participating components. Hence it's natural to study these systems in terms of different time evolutions.

In this paper we attempt a novel approach to model the MAS, with a respect of two characteristics, the concurrent aspect of the MAS and the heterogeneous of components (agents) perceived as a different time rates of agents plan execution.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 15–27, 2015.
DOI: 10.1007/978-3-319-19578-0_2

Models. For this aim, timed models are suitable. The durational actions timed automata (daTA) [2] are a form of timed automata [1], that admits a more natural representation of action durations and advocates carrying true concurrency (which are realistic assumptions for specifying in natural way MAS). It's based on the Maximality semantics [4]. Maximality semantics has been proved necessary and sufficient for carrying both the refinement process and action durations. The daTA model has been defined and a nice characterization of the model was presented in [2] and [7]. So the concurrency aspect of MAS is modeled by the timed automata with action durations.

The daTA model assumes a "global clock" semantics, i.e., all clocks advance simultaneously and at the same rate (and there is a common initial instant). All possible executions of daTA are then represented by an infinite transition system where, for any given state, the system may evolve in two possible ways: either it executes an action or it delays with a given amount of time the potential execution. The decidability of the daTA has been proved using the so-called region graph construction [1]. In this paper we are concerned by the coordination problem in MAS. Mainly, this consists in maintaining the synchronization of the agent plans with respect to some objective in a consistent timed context. We consider real time application wherein the agent plans refers to timed actions whose durations are known. Agents are assumed to be able to communicate via reliable materials, in order to achieve some plan called coordinated plan.

The paper contribution. We propose to model plan in a more attractive way using the standard algebraic language based on LOTOS, seeing plans as the execution of concurrent processes [3]. In fact, LOTOS specifications supplies modular concepts useful to describe some plan over several agents. Moreover, LOTOS-basic specifications are translated in daTA relative time rates (daTA-RT). daTA-RT compactly represents the possibly infinite behaviors of the coordinated plans, this model is concerned by the timing constraints defined over the MAS plans, taking into account relative time rates that distinguish the speeds of the agents in coordination. In this paper, we show how to build a (finite) region graph structure from a daTA-RT specification, for decidability assessment and proof. We extend by this proposal the result of a precedent work [12] over timed automata with relative time rates to model heterogeneity property.

Related work. In the literature, the two aspects, verification and time management works are generally focused on how to consider clocks and time for example see [13]. In the main cases, clocks are synchronized and used by all processes (read and reset). Or clocks can drift by a certain amount of time Δ, particularly as long as the processes do not communicate (via synchronizing actions). In [14], distributed systems are modeled by means of network of timed automata evolving in different rates. However, checking emptiness or universality turns out to be undecidable in the majority of cases. In [12], we investigate the decidability for verification assessments of real time systems with relative speed of clocks (what we call heterogeneity property). More precisely, under same hypothesis we answer this question positively over the timed automata with relative time rates.

Regarding designing MAS, several recent papers focus on the framework based on refinement paradigm. In [6], the authors propose a formal modeling of critical MAS that aims to derive a secure system implementation. The approach is based on Event B language. They are interesting by modeling fault tolerant MAS. In the same context of Event B specification language, authors in [8] propose a formal approach for self-organizing MAS. [9] Addresses a top-down approach for MAS protocol description using Finite State Automata (FSA) and multi-Role Interaction (mRI) abstraction. We don't forget the work [11] in which MAS are specified by AgLOTOS. This latter language captures communication of processes (i.e. agents) by message passing, in addition to classical features of concurrent processes. In our case the basic LOTOS is sufficient for what we intend to present.

Paper outlines. In Section 2, the specification of coordinated MAS plans is presented, based on LOTOS concepts, the relative time rates are explained and our running example is built. In Section 3, the daTA model is recalled, its extension to relative time is defined and an informal meaner of generating daTA structure from LOTOS expression is shown. The main contribution of this paper is proposed in Section 4, where the semantic of daTA is presented and formally defined. In all the paper the same example is used to clarify concepts and applications. The end section concerns conclusion and some immediate perspectives.

2 Coordinated Mas Specification

In this paper, a multi agent system is a tuple $\mathrm{MAS} = (\mathrm{Ag}, \mathrm{Plan}, \mathrm{Act}, \tau, \gamma)$ where Ag is set of agents; Plan is a set of agent plans, called coordinated plans since some of them are realized by several agents, Act is the set of actions mentioned in these plans. $\tau = (\tau_p)_{p\in \mathrm{Ag}}$ is a rate mapping associating a relative time rate with each agent characterizing the speed of the agent to execute their actions, γ is a duration mapping assigning a global duration value to each action of Act, evaluated in a number of execution cycles (called time units). In the following, we describe the specification of coordinated plans as an extension of the language Basic LOTOS, precising which parts of the plans are dedicated to each agent.

The relative time model. In the MAS systems, agents are assumed to have a notion of clocks to achieve their actions. Since agents can have different speeds, we assume that the speeds are relative according to a global time-scale (denoted absolute time), thus the duration performance of some action can be more or less important, depending on the agent considered to execute the action.

As example, Fig.1 shows an action a of duration 2, which requires 2 times more to be achieved by the agent p. According to any global time t, the time rate $\tau_q = 2t$ and $\tau_p = t$, in such a way that at any time $\frac{\tau_p}{\tau_q} \in Z$.

Coordinated plan specification. The plan of an agent p is specified by an agent expression E_p describing the actions to be executed for achieving the plan. The execution is assumed to be controlled by the agent p, however E_p can be composed of

sub-expressions whose execution can be performed by other agents. An agent expression inherits from the syntax of (Basic) LOTOS [3] as follows: $P ::= E_p$ and $E_p ::= exit \mid stop \mid a; E_p \ (a \in Act) \mid E_q \sim E_r$.

Where $q, r \in Ag$ and $\sim \in \{|||, ||, |[L]|, [], \gg, [>\}$ is a LOTOS operator.

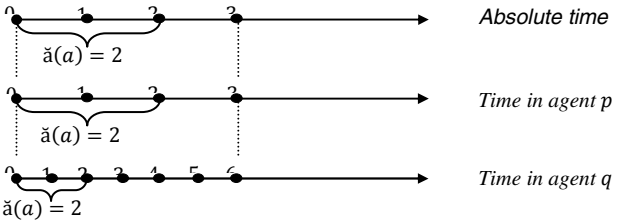


Fig. 1. The action a of global duration 2, is achieved two times faster by the agent q than by agent p

The elementary expression stop specifies a plan behavior without possible evolution and exit represents the successful termination of some plan. In the syntax, any operator $\sim$ as in standard LOTOS: $E_q[]E_r$ specifes a non-deterministic choice, $E_q \gg E_r$ a sequential composition and $E_q[> E_r$ the interruption of the left hand side part by the right one. The LOTOS parallel composition, denoted by $E_q|[L]|E_r$, can model both synchronous composition when $(L = Act)$, denoted by $E_q||E_r$, and asynchronous composition, $E_q|||E_r$ when $(L = \emptyset)$. In fact, the LOTOS language exhibits a rich expressivity such that the sequential executions of plans appear to be only a particular case.

Our running example. The example concerns two trucks A and B such that A initially placed in the location l_1 must pick up the load in the location l_2 and delivers it to the location l_4. As the load is initially placed in l_3, B initially placed in the location l_2, proposes to get the load from l_3 in such a way that A can meet it in l_2 and take the load. The problem for A and B is to meet them in minimum time, in case they start at the same time. In order to coordinate, each truck is equipped with a software agent able to discuss and synchronize with the other agents in the system. Both agents A and B refer to the following coordinated plan:

$P ::= E_A|[meet]|E_B$; with $E_A ::= move_A(l_2); meet; move_A(l_4); exit$ and
$E_B ::= move_B(l_3); get\_load_B; move_B(l_2); meet; exit.$

Moreover, duration of actions are given by the respective learnt experiences of A and B about transportations. For the simplicity of the example, all durations of actions are assumed equal to 1 time unit.

3 Timed Automata with Action Durations and Relative Time Rates for Coordination Plans

To model duration of actions, every edge of the automaton is annotated by constraints on clocks which implicitly enclose them, of course those that are already started. A single clock is reset on every edge. When clock is reset it corresponds to the beginning of event. The termination of action will be captured by information (temporal formulas) on locations of the automaton, precisely on the destination location. In fact, the duration of an action is either in the constraint of the following edge, if there is dependence between the successive actions, otherwise it is in the next locations and that means: action is not over yet. This elegant way to capture the durations is the effect of the maximality semantics. An example of a daTA A is shown in Fig. 2. The automaton consists of three localities l_0, l_1, l_2 and two clocks x, y. A transition from l_0 to l_1 represents the start of action a (indicating the beginning of its execution), the transition from l_1 to l_2 is labeled by b.

Assuming a time granularity of seconds, the automaton A starts in locality l_0. As soon as the value of y is less than or equal to 4, the automaton can make an a transition to l_1 and reset the clock x to 0. On the locality l_1 the temporal formula $\{x \geq 2\}$ represents information about the duration of the action a (it is important to differentiate it from invariant in timed automata). When x is at most 2 and is at least 5, transition to l_2 can be started (b executed) and y is reset. In the same logic the temporal formula $\{y \geq 7\}$ represents duration of the action b.

Preliminary. In the following we consider $R_{\geq 0}$ a set of nonnegative real numbers. Clocks are real variables take values from $R_{\geq 0}$. Let H be a set of clocks, a clock valuation over H is a function that assigns a nonnegative real number to every clock. V_H is the set of total valuation functions from H to $R_{\geq 0}$. A valuation is noted $v \in V_H$, and for $d \in R_{\geq 0}$, $v + d$ maps every clock x to $v(x) + d$. For $\lambda \subseteq H$, the valuation $v[\lambda := 0]$ is defined by: $(v[\lambda := 0])(x) = 0$ if $x \in \lambda, v(x)$ otherwise.

The set $C(X)$ of clock constraints C is defined by the grammar:

$C ::= \text{true} \mid \text{false} \mid x \sim c \mid C \wedge C$, where $x \in X$, $c \in \mathbb{N}$ and $\sim \in \{<, >, =, \leq, \geq\}$. We write $v \vDash C$ when the valuation v satisfies a clock constraint C over X iff C evaluates to true according to the values given by v.

We also use a subset of constraints where only the atomic form of clocks comparison is allowed. This set is defined by $C_d(H)$ by the grammar: $C ::= x \geq c$, where $x \in X$ and $c \in \mathbb{N}$. This subset represents condition duration over the set of actions noted by Act.

Definition 1 (daTA). A Durational Actions Timed Automaton daTA $\mathcal{A}$ is a tuple (S, s_0, H, T, L_S) over Act, where: S is a finite set of locations. $s_0 \in S$ is an initial location. H is a finite set of clocks. $T \subseteq S \times C(H) \times Act \times H \times S$ is a finite set of edges. An edge $e = (s, g, a, x, s') \in T$ $(s \xrightarrow{G,a,x} s')$ represents an edge from location s to s′ that launches the execution of action a whenever guard g becomes true.

$L_S: S \to 2^{C_d(H)}$ is a maximality function which decorates each location by a set of timed formulas named action durations. These formulas indicate the status of action execution at the corresponding state. $L_S(s_0) = \emptyset$ means that no action is yet started.

3.1 Timed Automata with Action Durations and Relative Speed Clocks Model

In this section we define a daTA with relative speed of clocks for modeling MAS; this is what we designate as the relative time rates in the global system.

Definition 2 (daTA with relative time rates). A daTA with relative time rates (daTA-RT) over the set of agents Ag is a structure $A = (\mathcal{A}, \pi)$ where $\mathcal{A} = (S, s_0, H, T, L_S)$ is a daTA and π is a mapping from H to Ag such that, for each $p \in Ag$, we have $\pi^{-1}(p) \subseteq H$.

Note that each clock evolves at the same rate in a particular agent (as the time evolves). This clock is then said to belong to that agent, and the mapping π (owner map) describes this in the above definition. We suppose that, in daTA with relative time rates, all clocks in H evolve at relative rates. Each rate characterizes the speed of an agent p. It depends on some absolute time given by the function $\tau_p: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\tau_p(0) = 0$ and $\tau_p(t)$ returns the local time in each agent $p \in Ag$ for the instant t of absolute time. Moreover, τ is a tuple of local time rate functions such that $\tau = (\tau_p)_{p \in Ag}$. The function τ_p is the p local time rate. For a time value t, the mapping function $\tau: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}^{Ag}$ assigns the tuple $(\tau_p(t))_{p \in Ag}$ to $\tau(t)$.

3.2 From Basic LOTOS with Action Durations to daTA with Relative Time Rates

In this section, we informally present the manner of generating in an operational way a daTA-RT starting from a Basic LOTOS specification with action durations. The approach is very close with the one of [4] for the generation of MLTS (Maximality based labeled transition system). In our context it's viewed as an unfolding operation of the behavior expression to state transition structure. Take again the behavior of the

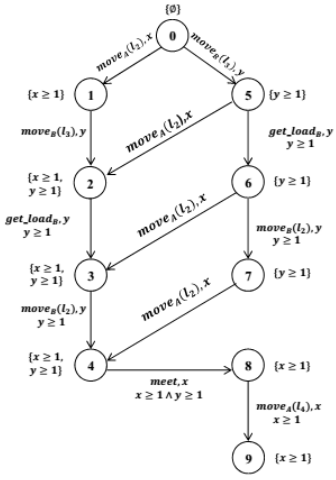


Fig. 3. A daTA-RT of the transport example (Our running example)

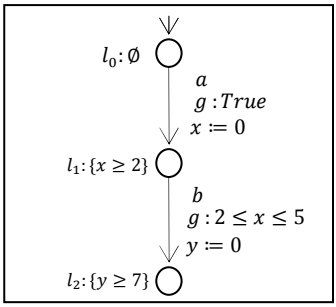


Fig. 2. A daTA

system described in the example of transport which is expressed as the parallel composition of two subsystems with synchronization on the action *meet*.

Recall that the behavior of system is specified by the following Basic LOTOS with action durations expression: $E_A||[meet]||E_B$. As already presented (in section 3.2), all states of the daTA-RT encapsulate a timed formula representing duration conditions of actions.

The generation of the daTA-RT's structure starts from the initial expressions, as a form of unfolding structure. To each configuration is associated a daTA-RT state as well as a behavior of sub-expression; (i.e. configurations corresponding daTA-RT's states) combines the expression to developed and duration conditions inherited from the previous step of unfolding, in the form $_{F}[E]$ where $F \in 2^{C_d(H)}$. In the initial state of the coordinated plan $P ::= E_A|[meet]|E_B$, no action is running, which explains why the duration conditions set is empty in the initial configuration. The initial configuration (State) of daTA-RT is $s_0 = {}_{\emptyset}[E_A|[meet]|E_B]$ where no action was launched yet. From this configuration, action $move_A(l_2)$ can be allowed. A clock x, is assigned to this occurrence of $move_A(l_2)$, it will be initialized to zero. Action $move_A(l_2)$ does not await the end of any other action to be able to comply, from where the guard of the transition is true.

$$\underbrace{{}_{\emptyset}[E_A]|[meet]|\,{}_{\emptyset}[E_B]}_{config_0} \xrightarrow{true,\ move_A(l_2),\ x} \underbrace{{}_{x\geq 1}[meet, move_A(l_4), exit]|[meet]|\,{}_{\emptyset}[E_B]}_{config_1}$$

The duration condition of the configuration $config_1$ corresponding to state s_1 is $\{x \geq 1\}$, which means that action $move_A(l_2)$ is running at this step of execution. From the configuration $config_1$ corresponding to state s_1, the only possible transition to construct is that corresponding to the launching of the action $move_B(l_3)$, from where the derivation

$$config_1 \xrightarrow{\emptyset,\ move_B(l_3),\ y} \underbrace{{}_{x\geq 1}[meet, move_A(l_4), exit]|[meet]|_{y\geq 1}[get\_load_B; move_B(l_2); meet; exit]}_{config_2}$$

With the same reasoning, we obtain $config_3$ and $config_4$, note the guard in this step is $\{y \geq 1\}$, this expresses dependence of loading B and the termination of moving B.

$$config_2 \xrightarrow{\{y\geq 1\},\ get_{load_B},\ y} \underbrace{{}_{x\geq 1}[meet, move_A(l_4), exit]|[meet]|_{y\geq 1}[move_B(l_2); meet; exit]}_{config_3}$$

The clock y is assigned to the action get_load_B because of its discharge at the time of get_load_B.

$$config_3 \xrightarrow{\{y\geq 1\},\ move_B(l_2),\ y} \underbrace{{}_{x\geq 1}[meet, move_A(l_4), exit]|[meet]|_{y\geq 1}[meet; exit]}_{config_4}$$

From the configuration $config_4$, action meet can comply only if the two actions $move_A(l_2)$ and $move_B(l_2)$ finished their execution, in other words, only if duration conditions belonging to the set $\{x \geq 1, y \geq 1\}$ are all satisfied, from where the set

$\{x \geq 1, y \geq 1\}$ corresponding to the condition $x \geq 1 \wedge y \geq 1$. The following transition becomes possible:

$$\text{config}_4 \xrightarrow{\{x\geq 1 \wedge y\geq 1\},\ \text{meet},\ x} \underbrace{{}_{x\geq 1}[\text{move}_A(l_4), \text{exit}]|[\text{meet}]|_{x\geq 1}[\text{exit}]}_{\text{config}_8}$$

For any action of synchronization, the assigned clock must evolve according to the slowest agent speed. In this example a clock x, which evolves according to the rate of truck A (τ_A), is assigned to the action meet. Thus, the truck B, which is the faster, must wait that truck A finished the synchronization action before going on. This mechanism ensures the synchronization on the end of each synchronized action. The configurations config_9 and config_{10} are obtained as follows:

$$\text{config}_8 \xrightarrow{\{x\geq 1\},\ \text{move}_A(l_4),\ x} \underbrace{{}_{x\geq 1}[\text{exit}]|[\text{meet}]|\ _{\emptyset}[\text{exit}]}_{\text{config}_9} \quad \text{and}$$

$$\text{config}_9 \xrightarrow{\{x\geq 1\},\ \delta,\ x} \underbrace{{}_{x\geq 0}[\text{stop}]|[\text{meet}]|_{x\geq 0}[\text{stop}]}_{\text{config}_{10}}$$

The same reasoning is applied in the following way to the other branch of the transition system where the action $\text{move}_B(l_3)$ begins its execution before the action $\text{move}_A(l_2)$:

$$config_0 \xrightarrow{\emptyset,\text{move}_B(l_3),y} \underbrace{{}_{\emptyset}[E_A]|[meet]|_{y\geq 1}[get\_load_B; move_B(l_2); meet; exit]}_{config_5}$$

The system cannot execute any action starting from configuration $config_{10}$, end of the unfolding operation of $E_A|[meet]|E_B$. The abstraction of this unfolding (transformation) is depicted by the daTA-RT structure in Fig. 3.

4 The Semantics of daTA-RT model

In the literature [1], an equivalence relation is proposed to aggregate states of the timed transition system (configurations) such that an equivalence class represents a configurations set. The equivalence classes of clock valuations are named clock regions. The design of Multi agents system becomes coherent if the components share a conjoint perception of time [5]. Therefore, it is important that the general perception is consistent. This perception takes its full dimension when calculating the semantics graph of the model. This motivates the redefinition of concepts like clock regions and region automaton. We will hereafter focus on the effect of the relative clock speeds.

4.1 Equivalence Classes of Clock Valuations

Let $A = ((S, L_S, s_0, H, T), \pi)$ be a daTA-RT over Act and a set of agents Ag.

Definition 3 ($slope_{xy}$). Let x (resp. y) be a clock that belongs to the agent p (resp. q) and evolves according to the rate function τ_p (resp. τ_q). We define slope_{xy} as the ratio of local-time rate functions τ_q and τ_p, noted $\text{slope}_{xy} = \frac{\tau_q}{\tau_p}$. (see Fig. 4).

Given a pair of clocks, x and y (within respectively agent p and q), their owner speeds will make them diverging from time reference at a certain speed. That is equal to the ratio between their owner rates. It represents the slope of the straight line in Fig. 4. As there are only finitely many clock constraints on clock x, we can determine the largest integer $c_x \in \mathbb{N}$ with which x is compared in some clock constraint (guard) of the daTA-RT A. In the remainder and for every pair of clocks x and y, the parameter slope_{xy} is assumed be an integer constant whatever the value of time t.

Definition 4 (equivalence relation ~). We define the equivalence relation ~ over the set of all clock valuations, $v \sim v'$ iff all the following conditions hold:

1. For all $x \in H$, either $\lfloor v(x) \rfloor$ and $\lfloor v'(x) \rfloor$ are the same, or both $v(x)$ and $v'(x)$ are greater than c_x.
2. For all $x, y \in H$ with $v(x) \leq c_x$, $v(y) \leq c_y$ and x (resp. y) evolves according to τ_p (resp. τ_q):

$c.\frac{1}{\text{slope}_{xy}} \leq v(x) \leq (c+1).\frac{1}{\text{slope}_{xy}}$ iff $c.\frac{1}{\text{slope}_{xy}} \leq v'(x) \leq (c+1).\frac{1}{\text{slope}_{xy}}$ for $c \in \mathbb{N}$ and $\text{fract}\left(\text{slope}_{xy} v(x)\right) \leq \text{fract}(v(y))$ iff $\text{fract}\left(\text{slope}_{xy} v'(x)\right) \leq \text{fract}(v'(y))$

An equivalence class of clock valuations induced by ~ is a clock region of A.

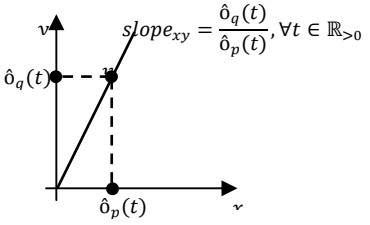


Fig. 4. Two clocks evolution with different speeds

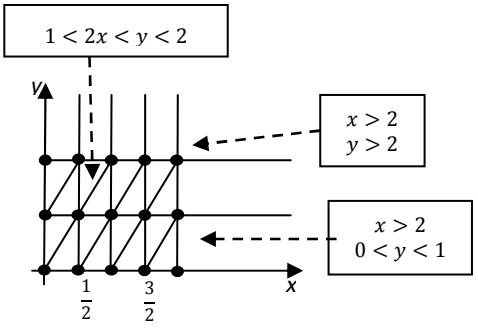


Fig. 5. Clock Regions Deducted by the Relation ~

4.2 The Representation of Clock Regions

Each equivalence class of clock valuations can be specified by a finite set of clock constraints it satisfies. The notation [v] represents the clock region to which v belongs.

Example. we consider two clocks x and y which evolve at different rates such that $\text{slope}_{xy} = 2$, $c_x = 2$ and $c_y = 2$. The clock regions obtained by the Definition 4 (equivalence relation) are depicted in Fig. 5. Thus, we have 15 corner points (e.g. $[x = 0{,}5 \wedge y = 2]$), 38 open line segments (e.g. $[0 < 2x = y < 1]$) and 23 open regions (e.g. $[0 < 2x < y < 1]$).

Definition 5 ($slope_{max}$). For each clock $x \in H$, we define $\text{slope}_{\max(x)}$ as the largest value of slope_{xy} for all $y \in H$.

Reconsider the example above: $\text{slope}_{\max(x)} = \max(\text{slope}_{xy}, \text{slope}_{xx}) = \max\left(\frac{2}{1}, 1\right) = 2$, $\text{slope}_{\max(y)} = \max(\text{slope}_{yx}, \text{slope}_{yy}) = \max\left(\frac{1}{2}, 1\right) = 1$. Note that if x is the fastest clock in H then $\text{slope}_{\max(x)} = \text{slope}_{xx} = 1$. $\frac{1}{\text{slope}_{\max(x)}}$ is the smallest amount of time in which x cannot stay in the same region.

In the example, the clock x changes the region each half unit of time corresponding to $\frac{1}{\text{slope}_{\max(x)}} = \frac{1}{2}$, when y do this change each one unit of time (except for regions represented by points). The representation of a clock region accords with the two following points:

For each clock x which evolves according to τ_p, there is one clock constraint taken from the set:

$$\{x = c \mid c = 0, \frac{1}{\text{slope}_{xy}}, 2\frac{1}{\text{slope}_{xy}}, \ldots, 1, 1 + \frac{1}{\text{slope}_{xy}}, 1 + 2\frac{1}{\text{slope}_{xy}}, \ldots, c_x \quad \text{for all } y \in H\}$$
$$\cup \left\{\wedge_{y \in H}\left(c - \frac{1}{\text{slope}_{xy}} < x < c\right) \mid c = \frac{1}{slope_{xy}}, 2\frac{1}{slope_{xy}}, \ldots, 1, 1 + \frac{1}{slope_{xy}}, 1 + 2\frac{1}{slope_{xy}}, \ldots, c_x. \forall y\right\}$$
$$\cup \{x > c_x\}. \quad (1)$$

for each pair of clocks x and y which evolve respectively according to τ_p and τ_q such that $c < x < c + \frac{1}{\text{slope}_{\max(x)}}$ and $d < y < d + \frac{1}{\text{slope}_{\max(y)}}$ appear in (1) for some c and d, whether $\text{slope}_{xy}(x - c)$ is less than, equal to or greater than $y - d$.

4.3 The Time-Successors of Clock Regions

In the following, we introduce the time-successor relation over clock regions. When time advances from any clock valuation v in the region α, we will reach all its time-successors α′. Formally, we say that α′ is a time-successor of the region α if there are v in α, v′ in α′, $t \in \mathbb{R}_{>0}$ such that $v' = v \oplus \tau(t)$, with $v \oplus \tau(t) = (v(x) + \tau_p(t))_{\pi^{-1}(x)=p}$.

For example, in Fig. 5 the five time-successors of the region $\alpha = [(1.5 < x < 2), (1 < y < 2x - 2)]$ are: itself, $[(x = 2), (1 < y < 2)]$, $[(x > 2), (1 < y < 2)]$, $[(x > 2), (y = 2)]$ and $[(x > 2), (y > 2)]$. These regions are those covered by a line drawn from any point in α parallel to the line $y = \text{slope}_{xy}x = 2x$ (in the upwards

direction). To compute each time-successor of a region α, we must give : (i). for every clock x, a constraint of the form $(x = c)$ or $\left(c < x < c + \frac{1}{\text{slope}_{\max(x)}}\right)$ or $(x > c_x)$ and (ii). for every pair x and y such that $(c < x < c + \frac{1}{\text{slope}_{\max(x)}})$ and $(d < y < d + \frac{1}{\text{slope}_{\max(y)}})$ appear in (i), the ordering relationship between $\text{slope}_{xy}(x - c)$ and $y - d$.

To compute the possible time-successors, three cases are distinguished: **First case:** Each clock x in the region α satisfies the constraint $(x > c_x)$, so α has only one time-successor, itself. This is the case of region $[(x > 2), (y > 2)]$ in Fig.5.

Second case: This case is considered when there is at least, in the region α, one clock x which satisfies the constraint $x = c$ for some $c \leq c_x$. The set H_0 contains all clocks appearing in similar constraint form as x. The clock region α will be changed immediately when the time advances, because the fractional part of each clock in H_0 becomes different from 0. The clock regions α and β have the same time-successors where β is specified by: A set of clock constraints which can be given as follows: For each clock $x \in H_0$: (i). If α satisfies $(x = c_x)$ then β satisfies $(x > c_x)$; (ii). If α satisfies $(x = c)$ then β satisfies $(c < x < c + \frac{1}{\text{slope}_{\max(x)}})$. For each clock $x \notin H_0$, the clock constraint in α remains the same in β. The ordering relationship between $\text{slope}_{xy}(x - c)$ and $y - d$ of each pair of clocks x, y in α is the same as that in β, such that $x < c_x$ and $y < c_y$ hold in the region α.

For example, in Fig. 5 the time-successors of the region $[(x = 0), (0 < y < 1)]$ are the same as the time-successors of the region $[0 < 2x < y < 1]$.

Final case: If the first and the second case do not apply, then let H_0 be the set of clocks x for which the region α satisfies two constraints: $c < x < c + \frac{1}{slope_{\max(x)}}$ $and slope_{xy}(x - c) \geq y - d$.

For all clocks y for which the region α satisfies $d < y < d + \frac{1}{slope_{max(y)}}$ Thus, when time advances, clocks in H_0 take the values $c + \frac{1}{slope_{max\,(x)}}$. Therefore, the time-successors of the region α are α, β and all the time-successors of β which is specified by :

1- A set of clock constraints which can be given as follows: (a) For each clock $x \in H_0$, if α satisfies $(c < x < c + \frac{1}{\text{slope}_{\max(x)}})$ then β satisfies $(x = c + \frac{1}{\text{slope}_{\max(x)}})$; (b) For each clock $x \notin H_0$, the clock constraint in α remains the same in β.

2- For each pair of clocks x and y such that $(c < x < c + \frac{1}{\text{slope}_{\max(x)}})$ and $(d < y < d + \frac{1}{\text{slope}_{\max(y)}})$ appear in (1.b), the ordering relationship between $\text{slope}_{xy}(x - c)$ and $y - d$ in α remains the same in β.

For example, in Fig. 5 the time-successors of the region $[0 < 2x < y < 1]$ include itself, $[(0 < x < 0{,}5), (y = 1)]$ and all time-successors of $[(0 < x < 0{,}5), (y = 1)]$.

Algorithm (region automaton). Let $A = ((S, s_0, H, L_S, T), \pi)$ be a daTA-RT over the set of agents Ag. The region automaton $R(A)$ is an automaton over the alphabet Act such that: - The configurations of $R(\mathrm{A})$ are of the form $< s|\alpha >$ where s is a state of A and α is a clock region. The initial configuration is of the form $< s_0|[v_0] >$ where $v_0(x) = 0$ for every $x \in H$. - A transition of $R(\mathrm{A})$, from the configuration $< s|\alpha >$ to $< s'|\alpha' >$, is labeled by $a \in Act$ iff there is a transition (s, G, a, x, s') in T and a clock region α'' which satisfies, α'' is a time-successor of the region α, and $\alpha'' \vDash g$ and $\alpha' = [\{x\} \mapsto 0]\alpha''$.

5 Conclusion

In this paper we have proposed an extension of the timed automata with action durations model, it has a capacity of describing timed plans with action durations that can be shared in between coordinated agents. We claim that also agents can be heterogeneous, they can reasonably be (re)synchronized to start the execution of any coordinated plan and that they can behave under relative time rates. Taking benefit from the semantics of such model, we demonstrated how to build a finite (time) region graph. The model is illustrated by a simple but realistic use case, wherein the coordination of truck is required. Agents are currently reduced to having one clock; however, the extension to several is immediate. *The implementability investigation*: In its current version our proposal is able to handle timed maximality bisimulation of behaviors; however region graph is not used for implementing practical tools because of the complexity of size and algorithms. A zone graph (based on convex polyhedra called zones) was proposed as an alternative for efficient implementations [10]; we intend to complete this work in this direction particularly in the sense of the scalability domain. As perspectives we intend to explore the ways of real applications and the formal comparison with other famous specification models such as petri net and its various extensions.

References

1. Alur, R., Dill, D.: A theory of timed automata. Theoretical Computer Science, 183–235 (1994)
2. Kitouni, I., Hachichi, H., Bouaroudj, K., Saidouni, D.E.: Durational Actions Timed Automata: Determinization and Expressiveness. International Journal of Applied Information Systems (IJAIS) 4(2), 1–11 (2012); Published by Foundation of Computer Science, New York, USA
3. Bolognesi, T., Brinksma, E.: Introduction to the ISO specification language LOTOS. Computer Networks and ISDN Systems 14, 25–59 (1987)
4. Courtiat, J.P., Saidouni, D.E.: Relating Maximality-based Semantics to Action Refinement in Process Algebras. In: Proceedings of FORTE 1994, pp. 293–308. Chapman and Hall (1995)
5. Lenzen, C., Locher, T., Wattenhofer, R.: Tight bounds for clock synchronization. In: Proceedings of the 28th ACM Symposium on Principles of Distributed Computing, pp. 46–55 (2009)

6. Pereverzeva, I., Troubitsyna, E., Laibinis, L.: Formal Development of Critical Multi-Agent Systems: A Refinement Approach. In: IEEE European Dependable Computing Conference (EDCC), pp. 156–161 (2012)
7. Guellati, S., Kitouni, I., Matmat, R., Saidouni, D.-E.: Timed Automata with Action Durations - From Theory to Implementation. In: Dregvaite, G., Damasevicius, R. (eds.) ICIST 2014. CCIS, vol. 465, pp. 94–109. Springer, Heidelberg (2014)
8. Corchuelo, R., Arjona, J.L.: A top down approach for MAS protocol descriptions. In: Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 45–49 (2003)
9. Graja, Z., Migeon, F., Maurel, C., Gleizes, M.-P., Kacem, A.H.: A Stepwise Refinement based Development of Self-Organizing Multi-Agent Systems. In: Dalpiaz, F., Dix, J., van Riemsdijk, M.B. (eds.) EMAS 2014. LNCS, vol. 8758, pp. 40–57. Springer, Heidelberg (2014)
10. Bouyer, P., Laroussinie, F.: Model Checking Timed Automata. In: Modeling and Verification of Real-Time Systems, pp. 111–140. ISTE Ltd. John Wiley & Sons, Ltd. (2008)
11. Chaouche, A.-C., El Fallah Seghrouchni, A., Ilié, J.-M., Saïdouni, D.E.: A Higher-Order Agent Model with Contextual Planning Management for Ambient Systems. In: Kowalczyk, R., Nguyen, N.T. (eds.) TCCI XVI. LNCS, vol. 8780, pp. 146–169. Springer, Heidelberg (2014)
12. Layadi, S., Kitouni, I., Belala, N., Saidouni, D.E.: About Decidability of Dynamic Timed Automata with Relative Time Rates. Submitted in: IGI-Global International Journal of Embedded and Real-Time Communication Systems, IJERTCS (2014)
13. Dima, C., Lanotte, R.: Distributed time-asynchronous automata. In: Jones, C.B., Liu, Z., Woodcock, J. (eds.) ICTAC 2007. LNCS, vol. 4711, pp. 185–200. Springer, Heidelberg (2007)
14. Akshay, S., Bollig, B., Gastin, P., Mukund, M., Kumar, K.N.: Distributed Timed Automata with Independently Evolving Clocks. Fundamenta Informaticae 130(4), 377–407 (2014)

Computational Intelligence: Object Recognition and Authentification

A Novel Technique For Human Face Recognition Using Fractal Code and Bi-dimensional Subspace

Benouis Mohamed[1], Benkkadour Mohamed Kamel[2], Tlmesani Redwan[2], and Senouci Mohamed[1]

[1] Computer Science Department, University Of Oran, Algeria
mhbenouis@yahoo.com, msenouci@yahoo.fr
[2] Computer Science Department, UniversitySidi Bel Abbes, Sidi Bel Abbes, Algeria
[2] INTTIC, Oran, Algeria
kamel.live@com, rtlemsani@ito.dz

Abstract. Face recognition is considered as one of the best biometric methods used for human identification and verification; this is because of its unique features that differ from one person to another, and its importance in the security field. This paper proposes an algorithm for face recognition and classification using a system based on WPD, fractal codes and two-dimensional subspace for feature extraction, and Combined Learning Vector Quantization and PNN Classifier as Neural Network approach for classification. This paper presents a new approach for extracted features and face recognition .Fractal codes which are determined by a fractal encoding method are used as feature in this system. Fractal image compression is a relatively recent technique based on the representation of an image by a contractive transform for which the fixed point is close to the original image. Each fractal code consists of five parameters such as corresponding domain coordinates for each range block. Brightness offset and an affine transformation. The proposed approach is tested on ORL and FEI face databases. Experimental results on this database demonstrated the effectiveness of the proposed approach for face recognition with high accuracy compared with previous methods.

Keywords: Biometric · Face recognition · 2DPCA · 2DLDA · DWT · PNN · WPD · IFS · Fractal codes · LVQ

1 Introduction

The security of persons, goods or information is one of the major concerns of the modern societies. Face recognition is one of the most commonly used solutions to perform automatic identification of persons. However, automatic face recognition should consider several factors that contribute to the complexity of this task such as the occultation, changes in lighting, pose, expression and structural components (hair, beard, glasses, etc.)[1]. several techniques have been proposed in the past in order to solve face recognition problems. Each of them evidently has their strengths and weaknesses, which, in most of the cases, depend on the conditions of acquiring

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 31–42, 2015.
DOI: 10.1007/978-3-319-19578-0_3

information. Recently, several efforts and research in this domain have been done in order to increase the performance of the recognition, such as support vector machine (SVM),Markov hidden model (HMM), probabilistic methods (Bayesian networks) and artificial neural networks. This latter has attracted researchers because of its effectiveness in detection and classification of shapes, which has been adopted in new face recognition systems [2].

2 Face Recognition System

A face recognition system is a system used for the identification and verification of individuals, which checks if a person belongs to the system's database, and identifies him/her if this is the case.
The methods used in face recognition based on 2D images are divided into three categories: global, local and hybrid methods.

- Local or analytical facial features approaches. This type consists on applying transformations in specific locations of the image, most frequently around the features points (corners of the eyes, mouth, nose,). They therefore require a prior knowledge of the images...
- Global approaches use the entire surface of the face as a source of information without considering the local characteristics such as eyes, mouth, etc.
- Hybrid methods associate the advantages of global and local methods by combining the detection of geometrical characteristics (or structural) with the extraction of local appearance characteristics.

This article is organized as follows: Basic notions concerning Two-dimensional subspace, wavelet transform theory are provided in Section 2. Fractal codes features are presented in section 3. Feature vectors results from two-dimensional subspaces is applied to a Combined LVQ and PNN classifier are described in Section 4. Section 5 provides face recognition system based on PNN, LVQ, the experimental results and Comparison between the serval's types of features obtained using WPD, DWT, IFS, 2DPCA and 2DLDA. A comparison with other approaches is also done in section 6. Conclusion and future works are presented in Section 7.

2.1 Two-Dimensional Principal Component Approach Analysis (2DPCA)

Proposed by Yang in 2004 [4], 2DPCA is a method of feature extraction and dimensionality reduction based on Principal Component Analysis (PCA) that deals directly with face images as matrices without having to turn them into vectors like as the traditional global approach.

2.2 The Steps of Face Recognition by 2DPCA

Considering training set S of N face images, the idea of this technique is to project a matrix X of size (n×m) via a linear transformation like that:

$$Y_i = X.R_i \quad (1)$$

Where Y_i is the principal component vector of size (n 1), and $R_i$is the base projection vector of size (m 1). The optimal vector $R_i$of the projection is obtained by maximizing the total generalized variance criterion

$$J(R) = R^T.G_t.R \quad (2)$$

Where G_t is the covariance matrix of size (m m) given by:

$$G_t = \frac{1}{M}\sum_{j=1}^{M}\left(X_j - \overline{X}\right)^T\left(X_j - \overline{X}\right) \quad (3)$$

With X_j: The j[th]image of the training set

$\overline{X}$: The average image of all the images in the training set.

$$\bar{X} = \frac{1}{M}\sum_{j=1}^{M} X_j \quad (4)$$

In general, one optimal projection axis is not enough. We must select a set of projection axes like:

$$\{R_1, R_2, \ldots\ldots, R_d\} = arg\ max\, J(R) \quad (5)$$

$R_i^T.R_j = 0, i \neq j, i, j = 1, \ldots\ldots, d$

These axes are the eigenvectors of the covariance matrix corresponding to the largest "d" Eigenvalues. The extraction of characteristics of an image using 2DPCA is as follows

$$Y_k = X.R_k \quad ; \text{k=1 ... d} \quad (6)$$

Where $[R_1, R_2, \ldots\ldots, R_d]$ is the projection matrix and $[Y_1, Y_2, \ldots\ldots, Y_d]$ is the features matrix of the image X.

2.3 The 2DLDA Approach

In 2004, Li and Yuan [5] have proposed a new two-dimensional LDA approach. The main difference between 2DLDA and the classic LDA is in the data representation model. Classic LDA is based on the analysis of vectors, while the 2DLDA algorithm is based on the analysis of matrices.

2.4 Face Recognition Using 2D LDA

Let X is a vector of the n-dimensional unitary columns. The main idea of this approach is to project the random image matrix of size $(m \times n)$ on X by the following linear transformation:

$$Y_i = A_j X \quad (7)$$

Y: the m-dimensional feature vector of the projected image A.

Let us suppose L: class numbers.
M: The total number of training images
The training image is represented by a matrix $m \times nA_j (j = 1, \ldots \ldots M)$
$\bar{A}_i$ (i=1 …L): The mean of all classes
N_i: Number of samples in each class

The optimal vector projection is selected as a matrix with orthonormal columns that maximizes the ratio of the determinant of the dispersion matrix of the projected inter-class images to the determinant of the dispersion matrix of the projected intra-class images;

$$J_{FLD}(X_{opt}) = \arg\max_W \frac{|X^T S_b X|}{|X^T S_w X|} \tag{8}$$

$$P_b = \text{trace}\ (S_b)$$

$$P_W = \text{trace}\ (S_w)$$

The unitary vector X maximizing J(X) is called the optimal projection axis. The optimal projection is chosen when X_{OPT} maximizes the criterion, as the following equation:

$$X_{OPT} = argmax_X J(X) \tag{9}$$

If S_W is invertible, the solution of optimization is to solve the generalized eigenvalue problem.

$$S_b X_{opt} = \lambda S_W X_{opt} \tag{10}$$

Like that λ is the maximum Eigenvalues of $S_W^{-1} S_b$
In general, it is not enough to have only one optimal projection axis. We need to select a set of projection axes,$x_1, x_2, \ldots . x_d$ under the following constraints:

$$\{x_1, x_2, \ldots . x_d\} = argmax_X J(X) \tag{11}$$

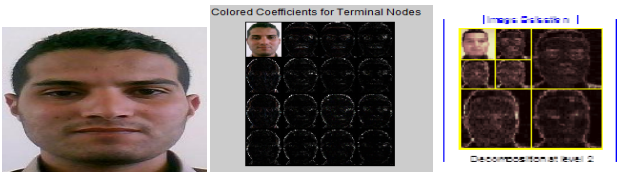
Indeed, the optimal projection axes$x_1, x_2, \ldots . x_d$ are orthonormal eigenvectors of $S_W^{-1} S_b$ corresponding to the best first "d" eigenvalues permitting to create a new projection matrix X, which is a matrix of size n × d :$X = [x_1, x_2, \ldots . x_d]$

We will use the 2DLDA optimal projection vectors $x_1, x_2, \ldots . x_d$ to extract the image features; we use the equation (08).

3 Discrete Wavelet Transform

Discrete wavelet transform (DWT) is a well-known signal processing field tool; it is widely used in feature extraction and compression and de-noising applications. The discrete wavelet transform has been used in various face recognition studies. The main advantage of the wavelet transform over the Fourier transform is the time-scale location. Mallat [8] shows that the DWT may be implemented using a filters bank including a low-pass filter (PB) and a high-pass filter (PH).

Discrete Wavelet Package Decomposition (D-WPD) is a wavelet transform where signal is passed through more filters that the Discrete Wavelet Transform (DWT). In the DWT, each level is calculated by passing only the previous approximation coefficients through low and high pass filters. However in the D-WPD, both the detail and approximation coefficients are decomposed [7] [8].



(a) (b) (c)

Fig. 1. Wavelet decomposition at different levels
(a) Original image
(b) 2-level wavelet decomposition using WPD
(c) 2-levels wavelet decomposition using DWT

4 Fractal Theory Codes

Fractal theory of iterated contractive transformation has been used in several areas of image processing and computer vision. In this method, similarity between different parts of an image is used for representing of an image by a set of contractive transforms on the space of images, for which the fixed point is close to the original image. This concept was first proposed by Barnsley [9], [10]. Jacquin was the first to publish an implementation of fractal image coding in [11]. Despite the number of researchers and the proposed methods, several factors can significantly affect face recognition performances, such as the pose, the presence/absence of structural components, facial expressions, occlusion, and illumination variations. Different image compression methods have been focused for a long time to reduce this massive information, but fractal image compression is a relatively recent technique based on representation of an image by contractive transforms, for which the fixed point is close to original image.

Suppose we are dealing with a 64*64 binary image in which each pixel can have on of 256 levels (ranging from black to white). Let R_1, R2,....., R_{256} be 4*4 non-overlapping sub-squares of the image (range blocks); and let D be the collection of all 8*8 pixel overlapping sub-squares of the image (Domain blocks) as depicted in Fig .2 The collection D contains 57*57=3249 squares. For each R block, search through all of D blocks a $D_i \in D$ which minimizes equation (12) .There 8 ways to map one square onto another. Each square can be rotated to 4 orientations or flipped and rotated into 4 other orientations as shown in Fig.2 having 8 different affine transformations means comparing 8*3249=25992 domain squares with each of the 256 range squares.

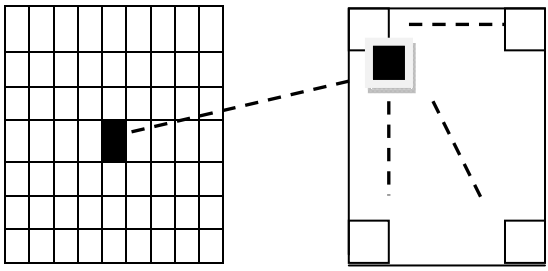


Fig. 2. One of the block mapping in partitioned function systems representation (IFS)

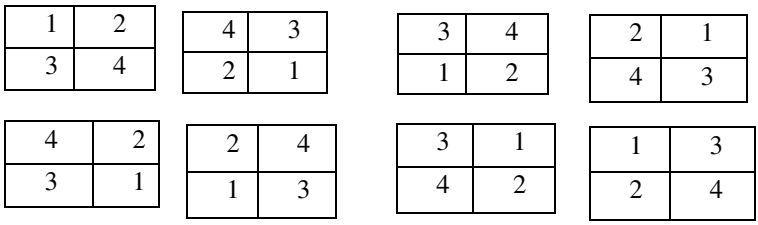


Fig. 3. Eight different affine transformations

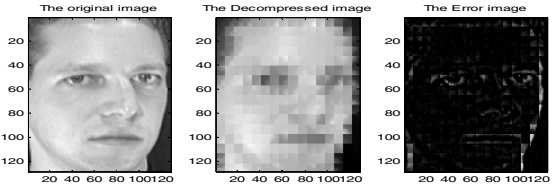
$$collage\ Error = min\|R_i - w(D_i)\|^2 \tag{12}$$

As mentioned before ,a D_i block has 4 times as many pixels as an R_i ,so we must either sub-sample (choose 1 from each 2*2 sub-square of D_i) or average the 2*2 sub-squares corresponding to each pixel of R when we minimize equation (12) .minimizing equation means two things .First it means finding a good choice for Di second, it means finding a good contrast and brightness setting S_i and O_i for W_i. In equation (13)

$$w_i\begin{bmatrix}x\\y\\z\end{bmatrix} = \begin{bmatrix}a_i & b_i & 0\\c_i & d_i & 0\\0 & 0 & s_i\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix} + \begin{bmatrix}c_i\\f_i\\o_i\end{bmatrix} \tag{13}$$

A choice of D_i , along with a corresponding S_i and $O_i$determines a map W_i. The type of image partitioning used for the range blocks can be so different. A wide variety of partitions have been investigated, the majority being composed of rectangular blocks. Different types of range block partitioning were described in[12].In this research we used the simplest possible range partition consists of the size square blocks, that is called fixed size square blocks (FSSB) partitioning. The procedure for finding a fractal model for a given image is called encoding; compression; or searching for a fractal image representation. After finding the best match ,fractal elements which of 6 real numbers (a, b, c, d, e, f) are selected as follows . (a, b, c, d) are (x, y) coordinates of the D block and its corresponding R block respectively .(e) is the index of affine

transformation that makes the best match.(it is a number between 1 and 8) ,(f) is the intensity is a number between 0 and 256.



(a) (b) (c)

Fig. 4. Decoding algorithm results (IFS)
(a) Original image
(b) Decoded image after 8 iteration for N=8
(c) The error image

In this paper, fractal code is introduced in order to extract the face features from the normalized face image based WPD. After fractal coding, where each domain is compared with all regions of the image, we obtain a set of transformations which can approximate the face image. Each transformation is represented by parameters of contrastS_i, brightness O_i, spatial coordinates of Range/Domain, and rotation W_i.The output of fractal code is the feature matrix with 2D-dimension used as a database of face which is applied two-dimensional subspace for reduction, discrimination and speed time.

5 Face Classification Using Neural Networks

Several studies have shown improved face recognition systems using a neural classification compared to classification based on Euclidean distance measure [14].

5.1 Probabilistic Neural Networks

The probability neural network is proposed by D. F. Specht for solving the problem of classification in 1988 [15]. The theoretical foundation is developed based on Bayes decision theory, and implemented in feed-forward network architecture.
PNN represent mathematically by the following expression

$$a = radbas(\|IW - x\|b) \tag{14}$$

$$y = compet(LW\alpha) \tag{15}$$

The structure PNN**:** The PNN architecture consists of two layers [15] [16]:

The first layer computes distances from input vector to the input weights (IW) and produces a vector whose elements indicate how close the input is to the IW.

The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities .Finally a compete transfer function on the output of the second layer picks up the maximum of these probabilities and produces a 1 for that class and a 0 for the other classes. The architecture for this system is shown above.

The probability of neural network with backs propagation networks in each hidden unit can approximate any continuous non linear function. In this paper, we use the Gaussian function as the activation function:

$$radbas = \exp\left[(-n^2)\right] \tag{16}$$

Finally, one or many larger values are chosen as the output unit that indicates these data points are in the same class via a competition transfer function from the output of summation unit [11], i.e.

$$compet(n) = e_i = [000\ 0_1 0 \ldots\ldots 0_i], n(I) = MAX(n)$$

5.2 Linear Vector Quantization (LVQ)

The vector quantization technique was originally evoked by Tuevo Kohonen in the mid 80's [15] [17]. Both Vector quantization network and self organizing maps are based on the Kohonen layer, which is capable of sorting items into appropriate categories of similar objects. Such kinds of networks find their application in classification and segmentation problems.

LVQ network comprises of three layers: Input layer, Competitive layer and Output layer [17]. The number of neurons in each layer depends on the input data and the class of the system. The input neurons are as many as the input matrix features of the training pattern, and the number of the output neurons is equal to the number of person's to which face patterns are classified. The number of hidden neurons is heuristic .In order to implement a face recognition system by our approach, we follow this methodology:

- stage pre-processing using technique WPD
- coding image using fractal code
- feature extraction using 2DPCA/2DLDA
- classification using LVQ and PNN network

6 Results and Discussion

In order to evaluate and test our approach described for face recognition system, we chose three databases: ORL, FEI [18] [19] and our database of our laboratory. All experiences were performed in Matlab installed on a laptop with a dual core processor T5870 with 2.03 GHz and 2 GB of RAM.

To evaluate the performance of our proposed approach, we chose two test databases: ORL and FEI. The global performance of algorithms tested on the FEI database is not as better as that of the ORL database. There are two main reasons:

- The image quality of the ORL database is better than that of the FEI database.

- The FEI database is more complex due to variations in the face details and head orientations.
- After a series of experiences, we chose the best values of parameters in order to fix the choice of Eigenvalues, which give a better recognition rate.

Adding Some Effects: It is wanted to test our system with and without added noisy in the two data base in order to evaluate robustness of these approaches namely 2DPCA, 2DLDA, DWT,WPD ,fractal codes combined by using two classifier LVQ and PNN.

Noise**:** Two types of noise are used in this simulation: the Saltand Pepper type noise with a noise density a=0.06 (Figure 5 (a)) and Gaussian noise with mean m=0, variance v=0.04 .Figure 5 illustrates these effects which are obtained as follows.

(a) Salt&pepper Noise **(b)** Gaussian Noise **(c)** Gaussian Noise m=0, v=0.01 m=0, v=0.04

(a) Salt&pepper Noise **(b)** Gaussian Noise **(c)** Gaussian Noise m=0, v=0.01 m=0, v=0.04

Fig. 5. Adding Noise (database face ORL &FEI)

The Pre-processing Stage: We proposed to add a preprocessing stage in order to improve our system's performance in speed by reducing the size and eliminating redundant information from the face images by the means of the DWT and WPD technique, and in other hand reduce the memory and compute of our neural network-training algorithm (PNN) and LVQ.

We performed face recognition analysis through WPD and DWT with various wavelet series: Daubechies, Gabor, Coiflets, Symlets and Gauss. In order to select a best wavelet to enhance a rate recognition.

The fractal code is used on WPD and DWT coefficients, derived from WPD to generate detailed high frequency features of animation which forms Feature set one. In order to have fractal feature vectors with the same length, the size of the face must be normalized (32 × 32). The normalized image is coded by 64 transformations using fractal code. Consequently, we obtained 320 fractal features as each transformation is coded on 5 parameters, as already explained in Section 3. Table 2 shows the performance of our system using fractal features for the two databases.

Feature Extraction Using 2DPCA/2DLDA: After reducing the dimensional of the face images using IFS. We used the 2DPCA and 2DLDA feature extraction approaches in order to extract the weight images (Features images in the new space) which must be converted into vectors before implementing the classifier network (LVQ and PNN).

Choice of the Number of Eigenvalues: Two dimensional methods do not escape this problem, and the choice of the appropriate number depends on the used method and faces database. In our experiences, we have selected the best eigenvalues corresponding to the best variance values (eigenvectors)

Selection Parameters and Architecture System Classifier

PNN :our neural network training algorithm used in system face recognition is not require many parameters compared other neural networks (MLP,BP,LVQ.etc),that only parameter that is needed for performance of the network is the smoothing parameter σ .Usually, the researchers need to try different σin a certain range to obtain one that can reach the optimum accuracy[16][17]. To get a higher recognition rate, we have made a series of experiments to choose the best smoothing parameter σ used in PNN. The probabilistic Neural Network used in our system is composed of two layers

Input Layer: The first layer is the input layer and the number of hidden unit is the number of independent variables and receives the input data (number of feature extraction for each approach used in this paper)

Output Layer: gives the number of faces used in the Database training (ex: ORL 200 person's).

LVQ: The changes of LVQ classifier parameters have a high effect on the classification results. In this paper, we found that the best learning rate increases the recognition rate of the system whereas the learning rate is a critical parameter that affected in the recognition process. We use a different number of learning rate (0.1, 0.2, 0.3, and 0.6) with 500 epochs and 80 hidden Neurons experiments.

Table 1. The recognition rate obtained by different methods on the database ORL with added noisy

Type of classifier	DWT-2DPCA	DWT-2DLDA	WPD-2DPCA	WPD-2DLDA
LVQ	93%	94%	**94%**	**96%**
PNN	94.8%	95%	**96%**	**98%**

Table 2. The recognition rate obtained by different methods on the database FEI with added noisy

Type of classifier	WPD-2DPCA	WPD-2DLDA	WPD/IFS-2DPCA	WPD/IFS-2DLDA
LVQ	90%	92 .8%	**95%**	**96%**
PNN	95%	96%	**99%**	**99%**

Table 3. The running time (s) obtained by different methods on the database FEI with added noisy

	DWT-2DPCA	**WPD-2DLDA**	**WPD&IFS-2DPCA**	**WPD&IFS-2DLDA**
PNN	1.20	1.25	**2.10**	**2.05**
LVQ	1.25	1.45	**2.08**	**1.98**

Discussion

After these series of experiments, we clearly see the superiority of the two-dimensional methods combined with a probabilistic neural classifier combining those of a LVQ classifier (table.1).

In table 2 ,we present the recognition rate obtained when using all fractal features ,and those reduced by the bi-dimensional subspace analysis .There is trade -off between encoding time and average of recognition rate because when N(domain range decreases ,size of features vector will increase so LVQ and PNN learns more details and its generalization ability become weak. As feature extraction is faster for N=8 and average of recognition rate is also fair so we encoded input faces with this R blocks size. The classification results for face is shown in table .2 for N=8.

We also note that the choice of optimal component and the choice smoothing parameter, which represents a better recognition, rate for methods, 2DPCA and 2DLDA and accuracy of classification PNN and LVQ.

In table 3, we present the running time obtained when using fractal codes .computational complexity of fractal encoding is the disadvantage of fractal features in our system which can be improving by adaptive search to speed-up fractal image compression.

7 Conclusion

In this paper, we propose an approach for face recognition based on the combination of two approaches, one used for the reduction of space and feature extractions in two dimensions and the other for classification and decision.

A hybrid approach is introduced in which, through the bi-dimensional subspace analysis, the most discriminating wavelet fractal features are extracted and used as the input of a neural network (LVQ, PNN). The performance of our method is both due to the fidelity of fractal coding for representing images, the WPD algorithm to speed up the features extraction step, and the 2DPCA and 2DLDA which highlights all discriminating features.

As a perspective, we propose to use this approach in an uncontrolled environment (video surveillance) based on video sequences (dynamic images) in order to make the task of face recognition more robust.

References

1. Jain, A.K. (ed.).: Handbook of Biometrics. Michigan State University, USA Patrick Flynn University of Notre Dame, USA Arun A. Ross West Virginia University, USA © Springer Science+Business Media, LLC (2008)
2. Pato, J.N., Millett, L.I. (eds.).: Biometric recognition challenges and opportunities Whither Biometrics Committee Computer Science and Telecommunications Board Division on Engineering and Physical Sciences Copyright by the National Academy of Sciences (2010)
3. Zhang, D., Zhou, Z.-H. (2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. Neuro Computing 69, 224–231 (2005)
4. Nguyen, N., Liu, W., Venkatesh, S.: Random Subspace Two-Dimensional PCA for Face Recognition. Department of Computing, Curtin University of Technology, WA 6845, Australia
5. Yang, J., Zhang, D.: Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition 26(1) (January 2004)
6. Noushath, S., Kumar, G.H., Shivakumara, P. (2D)LDA: An efficient approach for face recognition. Pattern Recognition 39(7), 1396–1400 (2006)
7. Mallat, S.: A theory of multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(7), 674–693 (1989)
8. Feng, G.C., Yuen, P.C., Dai, D.Q.: Human face recognition using PCA on wavelet subband. SPIE Journal of Electronic Imaging 9(2), 226–233 (2000)
9. Barnsley, M.: Fractals Everywhere. Academic Press, San Diego (1988)
10. Jacquin, A.E.: Fractal image coding: A review. Proc. of the IEEE 81, 1451–1465 (1993)
11. Jacquin, A.E.: A Fractal Theory of Iterated Markov Operators with Applications to Digital
12. Image Coding, PhD thesis, Georgia Tech, 1989. Y. Fisher, Fractal Image Compression: Theory and Application, Springer-Verlag Inc. (1995)
13. Ebrahimpour-Komleh, H.: Face recognition using fractal codes. In: Proceedings of International Conference on Image Processing 2001. IEEE, Thessaloniki (2001)
14. Nazish.: Face recognition using neural networks. Proc. IEEE INMIC 2001, 277–281 (2007)
15. Specht, D.F.: Probabilistic neural network and the polynomial adaline as complementary techniques for classification. IEEE Trans. Neural Networks 1(1), 111–121 (1990)
16. Neural network toolbox matlabUser's Guide COPYRIGHT 1992 - 2002 by The MathWorks, Inc.
17. Computational intelligence paradigms: theory & applications using MATLAB / S. Sumathi and Surekha Paneerselvam. 2010 by Taylor and Francis Group
18. ORL. The ORL face database at the AT&T (Olivetti) Research Laboratory (1992)
19. FEI. The FEI face database at the Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil (June 2005/March 2006)

Computational Intelligence: Image Processing

A New Rotation-Invariant Approach for Texture Analysis

Izem Hamouchene(✉) and Saliha Aouat

Artificial Intelligence Laboratory (LRIA), Computer science Department,
University of sciences and technology (USTHB), Algiers, Algeria
{ihamouchene,saouat}@usthb.dz

Abstract. Image processing and pattern recognition are one of the most important area of research in computer science. Recently, several studies have been made and efficient approaches have been proposed to provide efficient solutions to many real and industrial problems. Texture analysis is a fundamental field of image processing because all surfaces of objects are textured in nature. Thus, we proposed a new texture analysis method. In this paper, we proposed a novel texture analysis approach based on a recent feature extraction method called neighbor based binary pattern (NBP). The NBP method extract the local micro texture and is robust against rotation, which is a key problem in image processing. The proposed system extract two-reference NBP histograms from the texture in order to calculate a model of the texture. Finally, several models have been constructed to be able to recognize textures even after rotation. Textured images from Brodatz album database were used in the evaluation. Experimental studies have illustrated that the proposed system obtain very encouraging results robust to rotation compared to classical method.

Keywords: Rotation invariance · Texture analysis · Feature extraction · Neighbor based binary pattern

1 Introduction

Texture analysis is one of fundamental domain in image processing and computer vision. In today's world, automatic image processing without human intervention has become an active research area. In fact, there is not a strict definition of the texture, but the texture can be defined as a visual pattern composed of entities that have characteristic such as brightness, color, shape, size, etc. Texture is present in most of real life objects in nature. This makes it fundamental and essential to analyze images. Texture can be subdivided into coarse, micro, macro, regular, periodic, aperiodic, random and stochastic type [1].

Texture analysis has been presented by Haralick [5]. Different approaches have been developed structural, statistical and transformed based approach. These approaches have been applied in different, various and recent applications such as face recognition [2], Fingerprint matching [3] and image segmentation [4]. Textured images are analyzed by identifying the local and global properties of the images.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 45–53, 2015.
DOI: 10.1007/978-3-319-19578-0_4

One of the key problem of image analysis is rotation. Indeed, how recognize a researched texture even after rotation. The rotation invariant problem remains unsolved today. In this study, we proposed a new system robust against rotation and extract pertinent patterns of the texture.

This paper is organized as follows: The next section we explain the recent feature extraction method applied on the proposed system, the neighbor based binary pattern (NBP). In section 3 we present the architecture of our proposed system. Section 4 illustrates experimental results using the proposed system and the last section conclude the paper.

2 Neighbor Based Binary Pattern

The neighbor based binary pattern (NBP) is a very simple and efficient method to describe the texture. The NBP method was proposed for the first time by Izem et al. [6] [7]. This method was inspired by a famous feature extraction operator called Local Binary Pattern (LBP) [8] [9]. The important advantage of the LBP operator is its monotonic gray-scale transformation invariance [10] [11] [12] and its computational simplicity which makes it able to analyze an image in a very short time.

The idea of the NBP method is to consider one analysis window of 3x3 pixels. Each neighbor of the central pixel is thresholded by the next neighbor. Thus, if the central pixel is a noise it is not a problem because the value of the central pixel is not considered. In the other hand, if one neighbor is a noise, not all the pattern will be wrong but only 1 bit. . This minimizes the error rate of the recognition. So, each neighbor is encoded by the value 1 if its value is greater than the next neighbor is and 0 otherwise. The binary code is interpreted as a decimal number and represent the value of the central pixel in the NBP number. This process is illustrated in Fig. 1

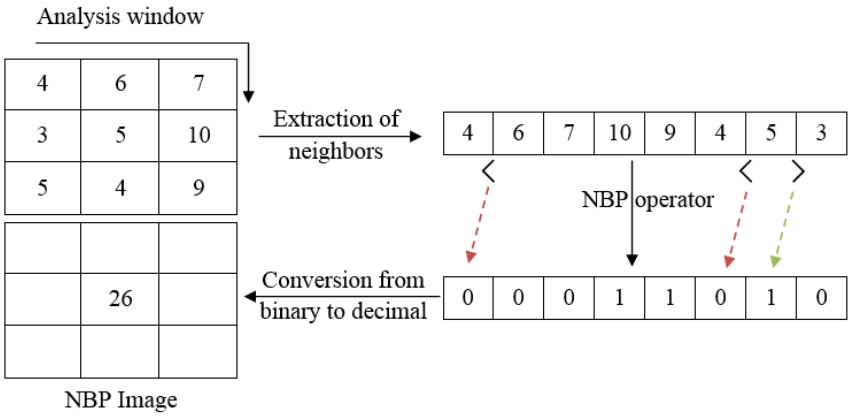


Fig. 1. Neighbor based Binary Pattern

Fig. 1. Illustrates the obtained NBP value using the NBP method. The first neighbor (value 4) is less than the second (value 6). Thus, the first neighbor is encoded by the value 0. After that, the obtained binary code is converted into a decimal number.

Because it is difficult to find a general parametric model for this distribution, the features of the obtained NBP image are approximated by a two dimensional discrete patterns histogram. This histogram is created to collect up the occurrences of each pattern. The obtained histogram is used to describe the texture as show in Fig. 2. Usually, the histograms are normalized.

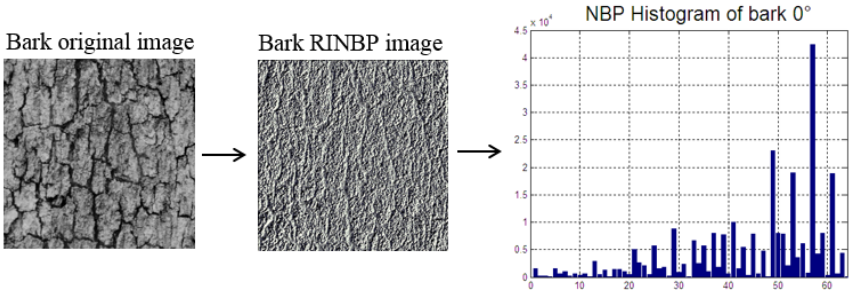


Fig. 2. Extraction of the NBP histogram

Fig. 2. Illustrates the extraction of the NBP histogram. We can notice from the Fig. 2 that a small rotation in the input image would cause a change in the output NBP code. Thus, if the patterns are extracted from the input image and this image is rotated by an angel θ, the extracted patterns will be different because the extraction starts always from the same point. That is the weakness of the classical LBP operator. In this work, we proposed a new system robust against the rotation problem.

3 Proposed system

In this section, we will explain the architecture of the proposed system. This system is robust against rotation and solve the rotation invariant problem. Some applications example will be illustrated and one illustration, which summarizes the proposed system, will be given.

The idea of the proposed system is to construct a model histogram of each texture. After that, compare the histogram of the researched texture with all model histograms to classify the texture. In order to solve the rotation problem, which is a fundamental problem on image processing and pattern recognition, we proposed to create a model histogram from each texture. First, two NBP histogram are extracted from two textured images, which have the same texture but different orientation. After that, a threshold histogram ($Hist_{th}$) is calculated based on the two NBP histograms. The threshold histogram contains the minimum and the maximum value of each bin of the two histograms. Indeed, the $Hist_{th}$ is the union of the two NBP histograms. This process is illustrated in Fig. 3.

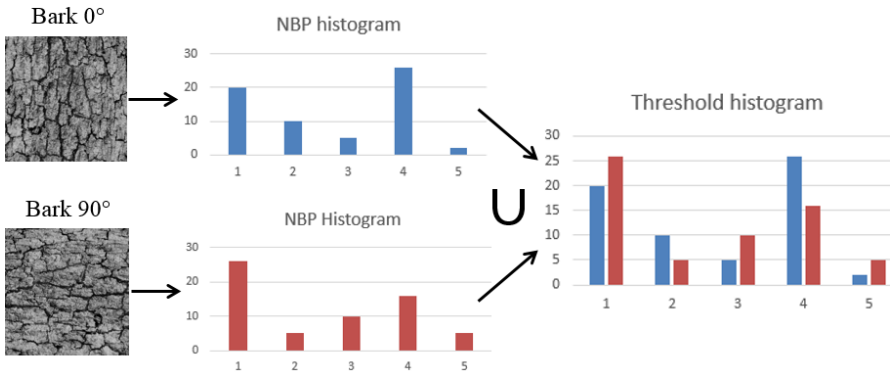


Fig. 3. Construction of the model histogram

Fig. 3. Illustrates the construction of the model histogram of the Bark texture. The threshold histogram of the Bark texture is constructed using the angles 0 and 90. Each bin of the threshold histogram, called model histogram, is an interval between the minimum and the maximum of the two NBP histogram of the two references images. Thus, a double threshold histogram is obtained and considered as a model of the texture.

In order to classify one query texture, a similarity distance is calculated between the NBP histogram of the query texture and the model histogram. The similarity vector is calculated following the formula 1.

$$\forall i \in [0,255], v(i) = \begin{cases} 1 \ if \ hist_q(i) \in hist_{th}(i) \\ 0 \ otherwise \end{cases} \quad (1)$$

Where his_{tq} is the NBP histogram of the researched texture. $Hist_{th}$ is the model histogram. Finally, a binary vector v is extracted where 1 means that the bin belongs to the interval of the model and 0 otherwise. The extraction of the similarity vector v is illustrated in Fig. 4.

Figure 4 illustrates the extraction of the similarity vector v. First, the NBP histogram is extracted from the query texture. After that, the intersection between the NBP histogram and the model histogram is encoded by 1, if the bin is in the double threshold, and 0 otherwise. The number of occurrence of the value 1 represents the similarity measure.

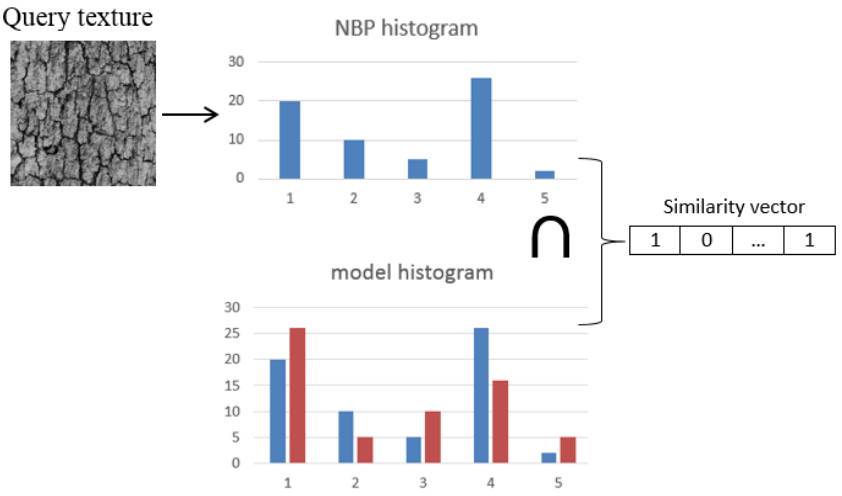


Fig. 4. Extraction of the similarity vector

4 Experimental results

In this section, the evaluation of the proposed system will be illustrated. In order to evaluate the performances of our proposed system, we used textured images from Brodatz album database [13]. Brodatz album is a famous benchmark for textured images. In the experimentation, we used twelve textured images (bark, brick, bubbles, grass, leather, pigskin, raffia, sand, straw, water, weave and wood) illustrated in Fig. 5.

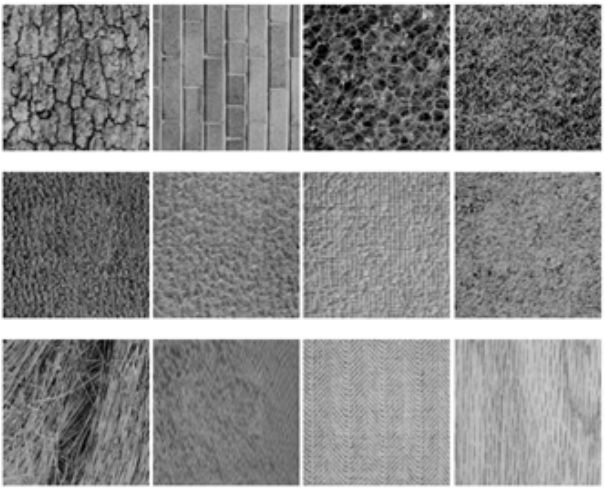


Fig. 5. Textured images from Brodatz album

Each image is digitized at seven different rotation angles: 0, 30, 60, 90, 120, 150, and 200 degree. The size of the images are 512x512 pixels with 256 gray levels as illustrated in Fig. 6, which contains a total of 84 images (12 different images with their 7 rotations).

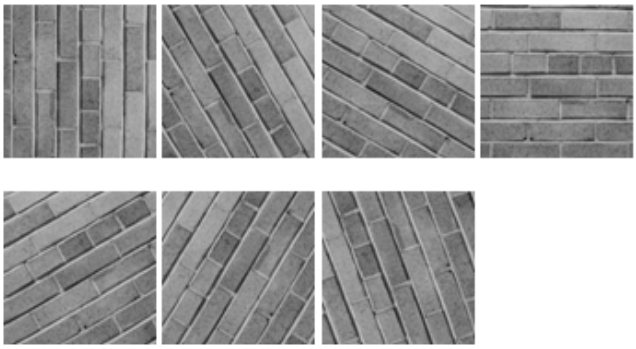


Fig. 6. Brick texture on seven orientation

In order to construct the model database of the system, model histograms are extracted from each texture. Thus, we obtain twelve model histograms. A classification process is applied to classify the query textures. First, the NBP histogram of the query texture is extracted. After that, the similarity measure is calculated between the NBP histogram of the researched texture and all model histograms of the system. Finally, the query texture is classified according to the most similar model.

In the evaluation part, all textured images of the database are considered as query texture. So, 84 test images (twenty images and its seven different orientation).

In order to compare the proposed system and the traditional system; the recognition averages of each texture are compared. The traditional texture classification system consider one textured image of each texture as reference. After that, the NBP histogram of the query texture is compared with all NBP histograms of the reference images and classified according to the most similar texture. The average recognition rate of each texture of the database (texture1 to texture 12) using the two recognition systems are illustrated in Table 1.

Table 1 illustrates the average of the recognition rate of each texture. We can notice that few textures are better recognized using the traditional method. However, most of all textures are well recognized using the proposed system and better classified. The traditional system consider only one orientation as reference. In fact, the weaknesses of the traditional system (rotation) are improved in our method with the model histogram (Double threshold histogram). This allows us to analyze the image with different orientations. This represents the strength of our method. The obtained comparison results are also illustrated in Fig. 7.

Table 1. Recognition rate of each texture

System	Traditional	Proposed
Bark	100,00	100,00
Brick	85,71	81,23
Bubbles	100,00	95,91
Grass	71,42	100,00
Leather	55,10	85,71
Pigskin	73,46	100,00
Raffia	51,02	100,00
Sand	100,00	100,00
Straw	42,85	95,26
Water	46,93	100,00
Weave	100,00	100,00
Wood	46,93	100,00
Recognition Rate	**67,19**	**89,09**

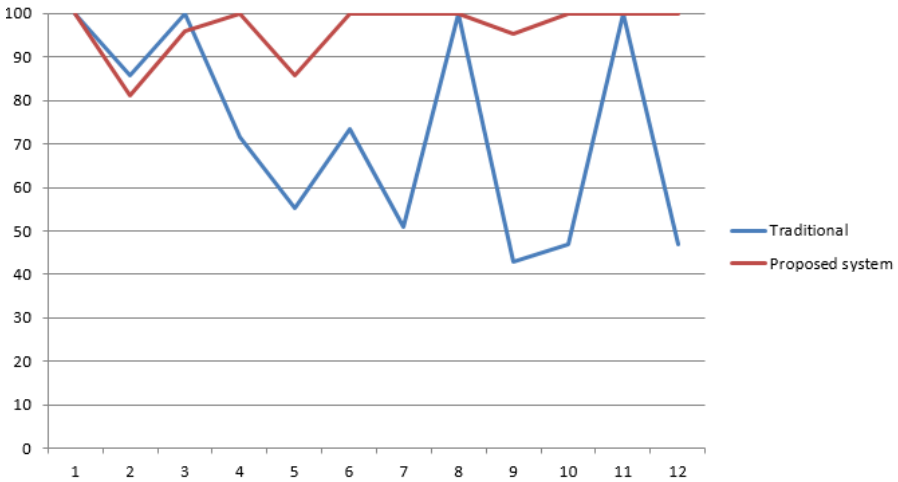


Fig. 7. Recognition rate using the traditional and proposed system

We can notice in the Fig. 7 that the traditional system (blue histogram) gives a lower performance compared to the proposed method (red histogram). Thus, we can see that the results given by our method are better than the classical method.

From the experiments and based on these obtained results, the global rate of the classical method is 67,19% and the proposed method is 89,09%. We can draw a conclusion that the proposed method, which is based on double threshold model histogram, is robust against rotation. Thus, the proposed system extract a robust model from the texture.

5 Conclusion

In this paper, we proposed a new rotation invariant system using the NBP method to extract models that describe the texture. First, the NBP method is applied on two textured images, which are in different orientation. After that, a double threshold histogram is calculated based on the two NBP histograms extracted from the two images and considered as model of the input texture. The model histogram is the union of the two NBP histograms. Finally, this process is applied on all textures of the database to extract models from each texture.

A similarity measure is calculated between the query texture and the model histograms. This measure is the intersection between the NBP histogram of the query texture and all model histograms. Each bin of the query texture histogram is encoded by 1 if its value is between the double threshold model histogram and 0 otherwise. Finally, a binary vector is extracted, which is the similarity vector. The query texture is classified according to the most similar vector. The obtained results show the efficiency of the proposed method compared to the traditional system. The robustness against rotation of the model histogram and the applied feature extraction method are the advantages of the proposed system.

In future works, we will combine other approaches to get more information from the texture like multi resolution methods. We will also improve the classification and the similarity measure to improve the recognition rate. We will also study the behavior and the robustness of our approach applied on real textured images.

References

1. Richards, W., Polit, A.: Texture matching. Kybernatic 16, 155–162 (1974)
2. Baohua, Y., Yuan, H., Jiuliang, C.: Combining Local Binary Pattern and Local Phase Quantization for Face Recognition. In: Biometrics and Security Technologies (ISBAST), pp. 51–53 (March 2012)
3. Jain, A.K., Ross, A., Prabhakar, S.: Fingerprint matching using minutiae and texture features. In: International Conference on Image Processing, vol. 3, pp. 282–285 (2001)
4. Hamouchene, I., Aouat, S., Lacheheb, H.: Texture Segmentation and Matching Using LBP Operator and GLCM Matrix. In: Chen, L., Kapoor, S., Bhatia, R. (eds.) Intelligent Systems for Science and Information. SCI, vol. 542, pp. 389–407. Springer, Heidelberg (2014)
5. Harlick, R.: Statistical and structural approaches to texture. Proc. of IEEE 67(5), 786–804 (1979)
6. Hamouchene, I., Aouat, S.: A New Texture Analysis Approach for Iris Recognition. In: AASRI Conference on Circuit and Signal Processing (CSP 2014), vol. 9, pp. 2–7 (2014)
7. Hamouchene, I., Aouat, S.: A cognitive approach for texture analysis using neighbors-based binary patterns. In: IEEE 13th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), August 18-20, pp. 94–99 (2014)
8. Ojala, T., Pietikäinen, M., Harwood, D.: A Comparative Study of Texture Measures with Classification Based on Feature Distributions. Pattern Recognition 29, 51–59 (1996)
9. Ojala, T., Pietikäinen, M.: Unsupervised Texture Segmentation Using Feature Distributions. Pattern Recognition 32, 477–486 (1999)

10. Guo, Z., Zhang, L., Zhang, D.: A Completed Modeling of Local Binary Pattern Operator for Texture Classification. IEEE Transactions on Image Processing 19(6), 1657–1663 (2010)
11. Xueming, Q., Xian-Sheng, H., Ping, C., Liangjun, K.: An effective local binary patterns texture descriptor with pyramid representation. Pattern Recognition 44(10-11), 2502–2515 (2011)
12. Baohua, Y., Yuan, H., Jiuliang, C.: Combining Local Binary Pattern and Local Phase Quantization for Face Recognition. In: Biometrics and Security Technologies (ISBAST), pp. 51–53 (March 2012)
13. Brodatz, P.: Textures: A Photographic Album for Artists and Designers. Dover Publications, New York (1966)

Multi-CPU/Multi-GPU Based Framework for Multimedia Processing

Sidi Ahmed Mahmoudi(✉) and Pierre Manneback

University of Mons, Faculty of Engineering, Computer science department
20, Place du Parc. Mons, Belgium
{Sidi.Mahmoudi,Pierre.Manneback}@umons.ac.be

Abstract. Image and video processing algorithms present a necessary tool for various domains related to computer vision such as medical applications, pattern recognition and real time video processing methods. The performance of these algorithms have been severely hampered by their high intensive computation since the new video standards, especially those in high definitions require more resources and memory to achieve their computations. In this paper, we propose a new framework for multimedia (single image, multiple images, multiple videos, video in real time) processing that exploits the full computing power of heterogeneous machines. This framework enables to select firstly the computing units (CPU or/and GPU) for processing, and secondly the methods to be applied depending on the type of media to process and the algorithm complexity. The framework exploits efficient scheduling strategies, and allows to reduce significantly data transfer times thanks to an efficient management of GPU memories and to the overlapping of data copies by kernels executions. Otherwise, the framework includes several GPU-based image and video primitive functions, such as silhouette extraction, corners detection, contours extraction, sparse and dense optical flow estimation. These primitives are exploited in different applications such as vertebra segmentation in X-ray and MR images, videos indexation, event detection and localization in multi-user scenarios. Experimental results have been obtained by applying the framework on different computer vision methods showing a global speedup ranging from 5 to 100, by comparison with sequential CPU implementations.

Keywords: GPU · Heterogeneous architectures · Image and video processing · Medical imaging · Motion tracking

1 Introduction

During the last years, the architecture of central processing units (CPUs) has so evolved that the number of integrated computing units has been multiplied. This evolution is reflected in both general (CPU) and graphic (GPU) processors which present a large number of computing units, their power has far exceeded the CPUs ones. In this context, image and video processing algorithms are well adapted for acceleration on the GPU by exploiting its processing units in parallel,

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 54–65, 2015.
DOI: 10.1007/978-3-319-19578-0_5

since they consist mainly of a common computation over many pixels. Several GPU computing approaches have recently been proposed. Although they present a great potential of GPU platform, hardly any is able to process high definition image and video efficiently and accordingly to the type of Medias (single image, multiple image, multiple videos and video in real time). Thus, there was a need to develop a framework capable of addressing the outlined problem.

In literature, one can categorize two types of related works based on the exploitation of parallel and heterogeneous platforms for multimedia processing: one related to image processing on GPU such as presented in [19], [12] which proposed CUDA[1] implementations of classic image processing and medical imaging algorithms. A performance evaluation of GPU-based image processing algorithms is presented in [15]. These implementations offered high improvement of performance thanks to the exploitation of the GPU's computing units in parallel. However, these accelerations are so reduced when processing image databases with different resolutions. Indeed, an efficient exploitation of parallel and heterogeneous (Multi-CPU/Multi-GPU) platforms is required with an effective management of both CPU and GPU memories. Moreover, the treatment of low-resolution images cannot exploit effectively the high power of GPUs since few computations will be launched. This implies an analysis of the spatial and temporal complexities of algorithms before their parallelization.

On the other hand, video processing algorithms require generally a real-time treatment. We may find several methods in this category, such as understanding human behavior, event detection, camera motion estimation, etc. These methods are generally based on motion tracking algorithms that can exploit several techniques such as optical flow estimation [6], block matching technique [20], and scale-invariant feature transform (SIFT) [9] descriptors. In this case also, several GPU implementations have been proposed for sparse [11] and dense [14] optical flow, Kanade-Lucas-Tomasi (KLT) feature tracker and SIFT feature extraction algorithm [17]. Despite their high speedups, none of the above-mentioned implementations can provide real-time processing of high definition videos. Our contribution consists on proposing a new framework that allows an effective and adapted processing of different type of Medias (single image, multiple images, multiple videos, video in real time) exploiting parallel and heterogeneous platforms. This framework offers:

1. Smart selection of resources (CPU or/and GPU) based on the estimated complexity and the type of media to process. In fact, additional computing units are exploited only in case of intensive and parallelizable tasks.
2. Several GPU-based image and video primitive functions ;
3. Efficient scheduling of tasks and management of GPU memories in case of Multi-CPU/Multi-GPU computations ;
4. Acceleration of several real-time image and video processing applications.

The remainder of the paper is organized as follows: section 2 presents our GPU-based image and video processing functions. The third section is devoted

[1] CUDA. https://developer.nvidia.com/cuda-zone

to describe the proposed framework for multimedia processing on parallel and heterogeneous platforms. Experimental results are given in section 4. Finally, conclusions and future works are discussed in the last section.

2 GPU-Based Primitive Functions

This section presents our image and video primitive functions that could be exploited by our framework for accelerating several computer vision methods.

2.1 Image Processing Primitive Functions

2.1.1 Noise Elimination we proposed the GPU implementation of noise elimination methods using the smoothing (or blurring) approach. The latter consists on applying a 2-D convolution operator to blur images and remove noise. We developed GPU version of linear, median and Gaussian filtering which represent the most used techniques for noise elimination. This GPU implementation consists of selecting the same number of CUDA threads as the number of image pixels. This allows for each CUDA thread to apply the multiplication of one pixel value with filter values. All the CUDA threads are launched in parallel. More details about this implementation are presented in [12].

2.1.2 Edges detection we proposed a GPU implementation of the recursive contours detection method using Deriche technique [3]. The noise truncature immunity and the reduced number of required operations make this method very efficient. Our GPU implementation of this method is described in [12], based on the parallelization of its four steps on GPU. Fig. 3(c) illustrates an example of edges detection.

2.1.3 Corners detection we developed the GPU implementation of Bouguets corners extraction method [2], based on Harris detector [5]. This method is efficient thanks to its invariance to rotation, scale, brightness, noise, etc. Our GPU implementation of this method is described in [16], based on parallelizing its four steps on GPU. Fig. 3(b) illustrates an example of corners detection.

Moreover, we have integrated the GPU module of the OpenCV [2] library that disposes of many GPU-based image processing algorithms such as FFT, Template Matching, histogram computation and equalization, etc.

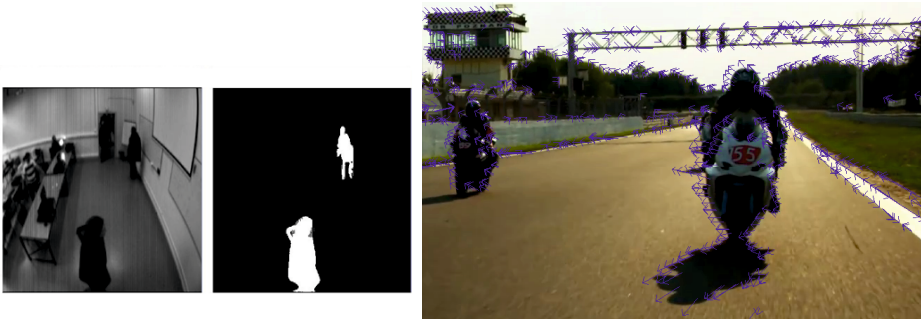
2.2 Video Processing Primitive Functions

2.2.1 Silhouette extraction the computation of difference between frames presents a simple and efficient method for detecting the silhouettes of moving objects, we propose a GPU implementation of this method using three steps.

[2] OpenCV GPU Module. www.opencv.org

First, we load the two first frames on GPU in order to compute the difference between them within CUDA in parallel. Once the first image displayed, we replace it by the next video frame in order to apply the same treatment. Fig. 1(a) presents the obtained result of silhouette extraction. This figure shows two silhouettes extracted, that present two moving persons. In order to improve the quality of results, a threshold of 200 was used for noise elimination.

2.2.2 Sparse optical flow estimation the sparse optical flow method consists of both features detection and tracking algorithms. The first one enables to detect features that are good to track, i.e. corners. To achieve this, we have exploited our corners extraction method (section 2.1.3). The second step enables to track the features previously detected using the optical flow method, which presents a distribution of apparent velocities of movement of brightness pattern in an image. It enables to compute the spatial displacements of images pixels based on the assumption of constant light hypothesis which supposes that the properties of consecutive images are similar in a small region. Our GPU implementation is detailed in [11]. Fig. 1(b) presents an example of sparse optical flow estimation using a Full HD video frame with characteristic points detected with the Harris corner detector and then tracked with the Lucas-Kanade method. Displacements are marked with arrows. Note that the arrows located on the static objects like trees or a building are there as a result of moving camera.



(a) GPU based silhouette extraction (b) GPU based sparse optical flow estimation

Fig. 1. GPU based video processing primitive functions

2.2.3 Dense optical flow estimation the GPU implementation of dense optical flow is based on the same process of sparse optical flow estimation. The only difference (compared to sparse) is that the tracking step is applied on all frames pixels. Thus, the number of selected CUDA threads is equal to the number of images pixels which requires more computation.

Notice that the image processing primitive functions have been adapted for treating videos also. Moreover, we have integrated the GPU based video processing algorithms of the OpenCV library such as frames interpolation, MOG (Mixture Of Gaussian) model, morphological operations, etc.

3 The Proposed Framework

The presented results and tests within sections 3 and 4 were run with Linux 64 bits on the following hardware:

- CPU: Intel Core (TM) i7, 980 3.33GHz, RAM : 8GB;
- GPU: 4 x NVIDIA GeForce GTX 580, RAM : 1.5GB.

The GPU-based primitive functions are exploited within our framework for processing different types of Medias: single image, multiple images, multiple videos and video in real time. The framework allows to select in an efficient way the adapted resources (CPU or/and GPU) in order to reduce the computation times with an optimal exploitation of computing units.

3.1 Single Image Processing on GPU

This kind of methods is applied on single images, which are displayed on screen at the end of processing. These algorithms are well adapted for GPU parallelization since they consist on common computations over many pixels. However, the use of graphics processing units offers high acceleration when processing high resolution images only. Indeed, performance can be either reduced with GPUs when treating low resolution images since we cannot benefit enough from the GPU. Therefore, we propose a treatment based on the estimated complexity of algorithms. The proposed treatment for single images is summarized in three steps: complexity estimation, resources selection, adapted processing.

3.1.1 Complexity estimation we propose to estimate the algorithm complexity $f_c$using the equation 1.

$$f_c = f \quad \times \quad comp\_pix \quad \times \quad size \tag{1}$$

where :

1. **f (Parallel fraction) :** Amdahl's law [4] proposed an estimation of the theoretical speedup using N processors. This law supposes that f is the part of program that can be parallelized and (1-f) is the part that can't be made in parallel (data transfers, dependent tasks, etc.). Indeed, high values of f can provide better performance and vice versa.

2. **comp_pix (computation per-image):** graphic processors enable to accelerate image processing algorithms thanks to the exploitation of the GPU's computing units in parallel. These accelerations become more significant when we apply intensive treatments since the GPU is specialized for highly parallel computation. The number of operations per pixel presents a relevant factor to estimate the computation intensity.
3. **size :** represents the resolution of input image.

3.1.2 Resources selection based on the estimated complexity f_c, we can have a good guidance for selecting the adapted resource (CPU or GPU) for computation. In fact, we launched for execution several GPU classic image processing (edge detection, corners detection. . .) algorithms using different image resolutions. These experiments allowed to define the value of f_c from which the GPU starts offering better performance than the CPU. This value is called the threshold S. Once the threshold defined, we compare the estimated complexity f_c for each input algorithm with the threshold S.

If $f_c > S$, the treatment is applied on GPU, else the CPU is used for processing. Notice that within our above-mentioned materiel, we have obtained a threshold S of 800000, that correspond to an algorithm with these parameters:

1. parallel fraction: 0.5 ;
2. number of operations per pixel comp_pixel: 10 ;
3. image resolution: 400×400.

We note also that the threshold value can change with other material configurations, since the number of GPUs computing units and the size of memories is not the same. Therefore, we propose to compute the threshold at each changement of material.

3.1.3 Adapted processing after selecting the adapted resource, CPU treatments are launched in case of low intensive algorithms ($fc < S$). The OpenCV library is employed for this aim. Otherwise, in case of high intensive algorithms ($fc > S$), we apply GPU treatment with three steps:

1. **Loading of input images on GPU :** first, the input images are loaded on GPU memory.
2. **CUDA parallel processing :** before launching the parallel processing of the current frame, the number of GPU threads in the so called blocks and grid has to be defined, so that each thread can perform its processing on one or a group of pixels in parallel. This enables the program to process the image pixels in parallel. Note that the number of threads depends on the number of pixels. Once the number and the layout of threads is defined, different CUDA functions (kernels) are executed sequentially, but each of them in parallel using multiple CUDA threads.

3. **OpenGL Visualization :** the output image is directly visualized on screen through the video output of GPU. Therefore, we propose to exploit the graphic library OpenGL enabling fast visualization, since it works with buffers already existing on GPU.

3.2 Multi-CPU/Multi-GPU Based Processing of Multiple Images

In case of multiple images treatment, performance can be less improved for two reasons: the first one is the inability to visualize many output images using only one video output that requires a transfer of results from GPU to CPU memory. The second constraint is the high computation intensity due to treatment of large sets of images. In order to overcome these constraints, we propose an implementation exploiting both CPUs and GPUs that offers a faster solution for multiple images processing. This implementation is based on the executive support StarPU [1] which offers a runtime for heterogeneous multicore platforms. For more detail, we refer authors to [8]. The employed scheduling strategy has been improved by taking into account the complexity factor f_c described in section 3.1.1. Indeed, high intensive tasks have higher priority for GPU computation. The low intensive tasks will be affected with a low priority for GPU. This allows to maximize the exploitation of available resources. As result, the repartition of tasks depends mainly on their computational intensity.

3.3 Multi-CPU/Multi-GPU Based Processing of Multiple Videos

This kind in methods is applied on a group of video sequences in order to extract some significant features. The latter can be exploited in several applications such as similarity computation between videos, videos indexation and classification. The real time processing is not required in this case. The treatment of a set of videos can be presented by the treatment of a set of images since a video is always represented by a succession of frames. Therefore, we propose a Multi-CPU/Multi-GPU treatment for multiple videos as shown in section 3.2.

3.4 Real Time Videos Processing on Multiple GPUs

In this case, we propose to exploit GPUs only since the video frames should be processed in order. This excludes the possibility of using heterogeneous platforms, which defines an order based on the employed scheduling strategy. Our approach of video processing on single or multiple GPUs consists of three steps:

1. **GPUs selection :** the program, once launched, first detects the number of GPUs in the system, and initializes all of them. Then, the input image frame is first uploaded to each GPU. This frame is virtually divided into equally sized subframes along y dimension and once the image data is available, each GPU is responsible for treating its part of the frame (subframe).

2. **Multi-GPU computation :** in this step, each GPU can apply the required GPU treatment (exp. optical flow computation). The related algorithm can be selected from our GPU primitive functions, or introduced by the framework user. We note also that the number of CUDA threads depends on the number of pixels within each subframe.
3. **OpenGL visualization :** at the end of computations for each frame (the subframes). The results can be displayed on screen using the OpenGL graphics library that allows for fast visualization, as it can operate on the already existing buffers on GPU, and thus requires less data transfer between host and device memories. In case of Multi-GPU treatments, each GPU result (subframe) need to be copied to the GPU which is charged of displaying. This, however, is a fast operation since contiguous memory space is always transferred. Once the visualization of the current image is completed, the program goes back to the first step to load and process the next video frames.

Otherwise, the framework can be used for processing multiple videos simultaneously using multiple GPUs. Indeed, each video stream is loaded and processed with one GPU. At the end of computations for each GPU (actual frame), the result is copied to the GPU which is charged for displaying. Each GPU result is visualized in a separated window in the same screen. Fig. 3.4 summarizes our framework showing the selected resources for each type of media. The figure shows also the primitive functions that could be exploited within the framework for accelerating different computer vision examples that require intensive computations.

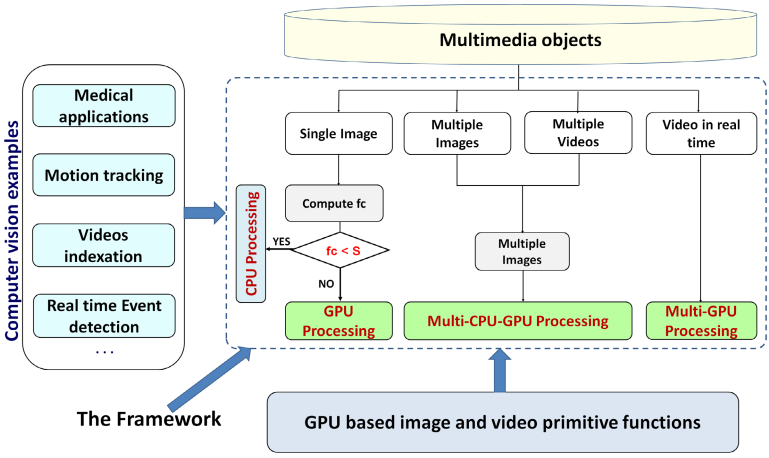


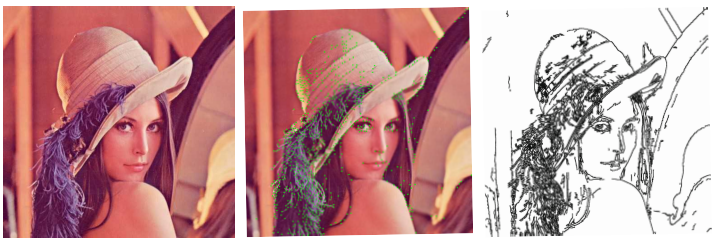
Fig. 2. Multi-CPU/Multi-GPU based Framework for Multimedia Processing

4 Experimental Results

The proposed framework has been exploited in several high intensive applications related to image and video processing such as image pre-processing, vertebra segmentation, videos indexation, event detection and localization.

4.1 CPU/GPU Based Image Pre-processing

Most of image processing methods apply a pre-processing step that allows to prepare the image for treatment. We can cite edges and corners detection methods which are so exploited for this aim. Based on our framework, we propose to accelerate these methods using CPU or GPU since the treatments are applied on single image. As presented in our framework, a complexity estimation is applied to select the convenient resource (CPU or GPU). Table 1 presents the selected resources and performance of corners and edges for different image resolutions. For each one, the complexity is evaluated using the above-mentioned metrics (section 3.1.1). The parallel fraction f presents the percentage of parallelizable computing part relative to total time, while the remaining part (1 - f) is presented by transfer (loading, visualization) operations. The computation per pixel is presented by the average of operations number between the steps of contours and corners detection. As result, the CPU is selected for treating low intensive methods, while the GPU is selected for high intensive ones. This allows to obtain fast results with a reduced energy consumption. In order to validate our results, we have calculated the ratio of acceleration (ACC) with GPU compared to CPU.



(a) Input image (b) Corners (c) Edges

Fig. 3. Edges and corners detection within our framework

Table 1. CPU/GPU based processing of single image processing (edges and corners detection), S = $8.0 * 10^5$

Images	**f**	$comp_{pix}$	f_c	$f_c > S$	**CPU/GPU ?**	**Acc**
256 × 256	0.55	6.1	$2.2 * 10^5$	No	**CPU**	**00.87** ↘
512 × 512	0.81	6.1	$1.3 * 10^6$	Yes	**GPU**	**05.88** ↗
1024 × 1024	0.86	6.1	$5.5 * 10^6$	Yes	**GPU**	**12.01** ↗
3936 × 3936	0.90	6.1	$8.5 * 10^7$	Yes	**GPU**	**19.85** ↗

As shown in Table 1, the GPU is selected only in case of methods that can benefit from the GPU's power. Otherwise, the CPU is selected. Fig. 3 presents an example of edges and corners detection within our framework.

4.2 Multi-CPU/Multi-GPU Based Vertebra Segmentation

The context of this application is the cervical vertebra mobility analysis on X-Ray or MR images. The main objective is to detect vertebra automatically. The computation time presents one of the most important requirements for this application. Based on our framework, we propose a hybrid implementation of the most intensive steps, which have been defined with our complexity factor f_c. Our solution for vertebra detection on Multi-CPU/Multi-GPU platforms is detailed in [8] for X-Ray images, and in [7] for MR images. Fig. 4(a) presents the results of vertebra detection in X-ray images, while Fig. 4(b) is related to present the detected vertebra in MR images. Notice that the use of heterogeneous platforms allowed to improve performance with a speedup of 30 × for vertebra detection within 200 high resolution (1472×1760) X-ray images, and a speedup of 98 × when detecting vertebra in a set of 200 MR images (1024 × 1024).

(a) Vertebra detection in X-ray images (b) Vertebra detection in MR images

Fig. 4. Vertebra detection in X-ray images

4.3 Multi-CPU/Multi-GPU Based Videos Indexation

The aim of this application is to provide a novel browsing environment for multimedia (images, videos) databases. It consists on computing similarities between videos sequences, based on extracting features of images (frames) composing videos [18]. The main disadvantage of this method is the high increase of computing time when enlarging videos sets and resolutions. Based on our framework, we propose a heterogeneous implementation of the most intensive step of features extraction in this application. This step, detected within our complexity estimation equation, is presented by the edge detection algorithm which provides relevant information for detecting motions areas. This implementation is detailed in [13] showing a total gain of 60% (3 min) compared to the total time of the application (about 5 min) treating 800 frames of a video sequence (1080x720).

4.4 Multi-GPU Based Event Detection and Localization in Real Time

This application is used for event detection and localization in real time. It consists of modeling normal behaviors, and then estimating the difference between the normal behavior model and the observed behaviors. These variations can be labeled as emergency events, and the deviations from examples of normal behavior are used to characterize abnormality. Once the event detected, we localize the areas in video frames where motion behavior is surprising compared to the rest of motion in the same frame. Based on our framework, we propose a Multi-GPU implementation of the most intensive steps of the application. The latter are also defined within the above-mentioned complexity factor f_c. This implementation is detailed in [10]. Notice that performed tests show that our application can turn in multi-user scenarios, and in real time even when processing high definition videos such as Full HD or 4K standards. Moreover, the scalability of our results is achieved thanks to the efficient exploitation of multiple graphic cards. A demonstration of GPU based features detection, features tracking, and event detection in crowd video is shown in this video sequence: https://www.youtube.com/watch?v=PwJRUTdQWg8..

5 Conclusion

We proposed in this paper a new framework that allows an adapted and effective exploitation of Multi-CPU/Multi-GPU platforms accordingly to the type of multimedia (single image, multiple images, multiple videos, video in real time) objects. The framework enables to select firstly the computing units (CPU or/and GPU) for processing, and secondly the methods to be applied depending on the type of media to process and the algorithm complexity. Experimental results showed different use case applications that have been improved thanks to our framework. Each application has been integrated in an adapted way for exploiting resources in order to reduce both computing time and energy consumption. As future work, we plan to improve our complexity estimation by taking into account more parameters such as tasks dependency, GPU generation, etc. we plan also to include primitive functions related to 3D image processing within our framework. The latter will be exploited for several medical imaging applications that could be applied larger sets of images and videos.

References

1. Augonnet, C., Thibault, S., Namyst, R., Wacrenier, P.-A.: StarPU: A Unified Platform for Task Scheduling on Heterogeneous Multicore Architectures. In: Sips, H., Epema, D., Lin, H.-X. (eds.) Euro-Par 2009. LNCS, vol. 5704, pp. 863–874. Springer, Heidelberg (2009)
2. Bouguet, J.Y.: Pyramidal Implementation of the Lucas Kanade Feature Tracker, Description of the algorithm. Intel Corporation Microprocessor Research Labs, 851–862 (2000)

3. Deriche, R., Blaszka, T.: Recovering and characterizing im-age features using an efficient model based approach. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, New York, USA, pp. 530–535 (1993)
4. Grama, A., Gupta, A., Karypis, G., Kumar, V.: Introduction to Parallel Computing, 2nd edn. Pearson Education Limited (2003)
5. Harris, C.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–152 (1988)
6. Horn, B.K.P., Schunk, B.G.: Determining Optical Flow. Artificial Intelligence 2, 185–203 (1981)
7. Larhmam, M.A., et al.: A portable multi-cpu/multi-gpu based vertebra localization in sagittal mr images. In: International Conference on Image Analysis and Recognition, ICIAR 2014, pp. 209–218 (2014)
8. Lecron, F., et al.: Heterogeneous computing for vertebra detection and segmentation in x-ray images. International Journal of Biomedical Imaging: Parallel Computation in Medical Imaging Applications 2011, 1–12 (2011)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) 60(2), 91–110 (2004)
10. Mahmoudi, S.A., et al.: Multi-gpu based event detection and localization using high definition videos. In: International Conference on Multimedia Computing and Systems (ICMCS), pp. 81–86 (2014)
11. Mahmoudi, S.A., Kierzynka, M., Manneback, P., Kurowski, K.: Real-time motion tracking using optical flow on multiple gpus. Bulletin of the Polish Academy of Sciences: Technical Sciences 62, 139–150 (2014)
12. Mahmoudi, S.A., Lecron, F., Manneback, P., Benjelloun, M., Mahmoudi, S.: GPU-Based Segmentation of Cervical Vertebra in X-Ray Images. In: IEEE International Conference on Cluster Computing HPCCE Workshop, pp. 1–8 (2010)
13. Mahmoudi, S.A., Manneback, P.: Efficient exploitation of heterogeneous platforms for images features extraction. In: 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 91–96 (2012)
14. Marzat, J., Dumortier, Y., Ducrot, A.: Real-time dense and accurate parallel optical flow using CUDA. In: Proceedings of WSCG, pp. 105–111 (2009)
15. Park, K., Nitin, S., Man, H.L.: Design and Performance Evaluation of Image Processing Algorithms on GPUs. IEEE Transactions on Parallel and Distributed Systems 28, 1–14 (2011)
16. Ricardo Possa, P., Mahmoudi, S.A., Harb, N., Valderrama, C., Manneback, P.: A multi-resolution fpga-based architecture for real-time edge and corner detection. IEEE Transactions on Computers 63, 2376–2388 (2014)
17. Sinha, S.N., Fram, J.-M., Pollefeys, M., Genc, Y.: Gpu-based video feature tracking and matching. In: EDGE, Workshop on Edge Computing Using New Commodity Architectures (2006)
18. Tardieu, D., al.: Video navigation tool: Application to browsing a database of dancers' performances. In: QPSR of the numediart research program, vol. 2(3), pp. 85–90 (2009)
19. Yang, Z., Zhu, Y., Pu, Y.: Parallel Image Processing Based on CUDA. In: International Conference on Computer Science and Software Engineering China, pp. 198–201 (2008)
20. Zhu, S., Ma, K.-K.: A new diamond search algorithm for fast block-matching motion estimation. IEEE Transactions on Image Processing 9(2), 287–290 (2000)

Full-Reference Image Quality Assessment Measure Based on Color Distortion

Zianou Ahmed Seghir[1(✉)] and Fella Hachouf[2]

[1] University Khenchela, Faculty. ST,
ICOSI Lab., BP 1252 El Houria, 40004 Khenchela, Algeria
[2] Laboratoire d'Automatique et de Robotique, Université Constantine1, Algeria
zianou_ahmed_seghir@yahoo.fr

Abstract. The purpose of this paper is to introduce a new method for image quality assessment (IQA). The method adopted here is assumed to be Full-reference measure. Color images that are corrupted with different kinds of distortions are assessed by applying a color distorted algorithm on each color component separately. This approach use especially *YIQ* color space in computation. Gradient operator was successfully introduced to compute gradient image from the luminance channel of images. In this paper, we propose an alternative technique to evaluate image quality. The main difference between the new proposed method and the gradient magnitude similarity deviation (GMSD) method is the usage of color component for the detection of distortion.

Experimental comparisons demonstrate the effectiveness of the proposed method.

Keywords: Gradient similarity · Quality assessment · Test image · Color distortion · Color space

1 Introduction

Over the past decade, image quality assessment methods based objective methods have grown significantly to tackle problems of image assessment. The challenge of these problems is to construct an algorithm that can automatically predict perceived quality of image.

There is no doubt that the subjective test is the most accurate measure for quality assessment because it reflects the true human perception. On the other hand, it is time consuming and expensive. There are three kinds of measures that are used for objective image quality assessment, full-reference (FR), reduced-reference (RR) and no-reference (NR). In this paper, the discussion is confined to FR metrics, where the reference images are available.

There has been extensive work on objective image quality assessment. The most popular method for full reference image quality assessment is the Structural Similarity Index [2] (*SSIM*). It contains three parts: Luminance Comparison, Contrast Comparison and Structure Comparison. However, it fails in measuring the badly blurred

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 66–77, 2015.
DOI: 10.1007/978-3-319-19578-0_6

images [3]. In [4], an approach based on edge-region information, distorted and displaced pixels (ERDDM) is developed. Initially, the test and reference images are divided into blocks of 11×11 pixels, and then distorted and displaced pixels are calculated which can be used to compute the global error. In [6], *DTex* metric is proposed with consideration of the texture masking effect and contrast sensitivity function. In [17], it was shown that the masking effect and the visibility threshold can be combined with structure, luminance and contrast comparison to create the image quality measure (gradient similarity measure (GSM)). Most Apparent Distortion (MAD) designed in [23, 24] yields two quality scores, i.e., visibility-weighted error and the differences in log-Gabor subbands statistics. The proposed measure in [13] applies phase congruency [15] to image quality measure. This measure differs in their correlations with the subjective quality and carrying out times. Gradient magnitude similarity deviation (GMSD) is proposed [14], where the pixel-wise gradient magnitude similarity (GMS) is used to capture image local quality, and the standard deviation of the overall GMS map is computed as the final image quality index.

The gradient images are sensitive to image distortions, whereas different local structures in a distorted image suffer different degrees of degradations. This motivates us to investigate the use of global variation of gradient based local quality map for overall image quality prediction. In fact, color deformation cannot be well differentiated by gradient. In addition, the gradient is computed from the luminance channel of images. Therefore, to make the image quality assessment measures own the ability to deal with color distortions, chrominance information should be taken into consideration.

The aim of this paper is to improve the GMSD to take color distortion in consideration. As a result, we use a proposed gradient operator and *YIQ* color space [1] to produce gradient image and color distortion from the reference and test images, respectively.

The rest of the paper is organized as follows. In Section 2, our proposed image quality measure is defined. In section 3, performance of the proposed method is compared with others measures using images with different types of distortion. We finish by the conclusion.

2 Proposed Method

Before introducing the proposed measure notion, some useful concepts must be visited. The reference and test images are represented by $Ref\ (M,N)$ and $Dis\ (M,N)$ respectively.

The proposed method uses gradient similarity and Color distortion to form map.

In addition, all variables used in the proposed method are defined next:

Ref: reference image.
Dis: test image.
$M \times N$: the image size.
G_1: gradient image of *Ref.*
G_2 : gradient image of *Dis.*
G_map: Gradient similarity map.

CFI_map and *CFQ_map* : chromatic features.
C_1, C_2: positive constants.
$GSCDM$: Gradient similarity based Color distortion measure.

2.1 Gradient Similarity

In order to reflect the differences between *Ref* and *Dis* at the local level, we compute image gradient of the reference and test images. Different operators are used to compute the image gradient, such as the Sobel operator [7], the Prewitt operator [7] and the Scharr operator [8], and in this paper a new gradient operator is proposed, which shows very favorable outcome. It defines as:

	Gx	*Gy*
Mask	$\begin{pmatrix} 4 & 0 & -4 \\ 3 & 0 & -3 \\ 4 & 0 & -4 \end{pmatrix}/11$	$\begin{pmatrix} 4 & 3 & 4 \\ 0 & 0 & 0 \\ -4 & -3 & -4 \end{pmatrix}/11$

This later consists of a pair of 3×3 convolution kernels and is used for detecting vertical and horizontal edges in images.

The partial derivatives *Gx* and *Gy* of an image are computed as:

$$G = \sqrt{Gx^2 + Gy^2} \tag{1}$$

Also, the gradient operators (G) of the reference and test images are computed. As a result, the G_2 and G_1 of the test and reference images are produced, respectively.

The gradient similarity is computed in proposed method and hence the Gradient map (***G_map***) is formed as

$$G\_map = \frac{2G_1.G_2 + C_1}{{G_1}^2 + {G_2}^2 + C_1} \tag{2}$$

2.2 Color Space Transformation

The color distortion cannot be differentiating by gradient. Hence, to make the image quality assessment measures possess the ability to deal with color distortions, special considerations are given to chrominance information. As a result, these formulas approximate the conversion between the *RGB* color space and *YIQ* [1]

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.144 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.528 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3}$$

Let I_1 (I_2) and Q_1 (Q_2) be the I and Q chromatic channels of the reference and distorted images respectively. Similar to the definitions of *CFI_map* and *CFQ_map*, the similarity between chromatic features is defined as follows:

$$CFI\_map = \frac{2I_1.I_2 + C_2}{{I_1}^2 + {I_2}^2 + C_2}$$
$$CFQ\_map = \frac{2Q_1.Q_2 + C_2}{{Q_1}^2 + {Q_2}^2 + C_2} \tag{4}$$

The similarity between the chrominance components (color distortion map) is simply defined as:

$$CD\_map = CFI\_map.CFQ\_map \tag{5}$$

2.3 Global Error

Finally, the gradient similarity based Color distortion map (***GSCD*_map**) is expressed as:

$$GSCD\_map = G\_map\,.\,CD\_map \tag{6}$$

The total gradient similarity based Color distortion measure (*GSCDM*) is defined as the standard deviation of the ***GSCD* map**:

$$GSCDM = \sqrt{\frac{1}{N.M}\sum_{p=1}^{M}\sum_{q=1}^{N}\left(\overline{GSCD} - GSCD\_map(p,q)\right)^2} \tag{7}$$

Where

$$\overline{GSCD} = \frac{1}{N.M}\sum_{p=1}^{M}\sum_{q=1}^{N} GSCD\_map(p,q) \tag{8}$$

Flowchart depicting computation of the proposed measure is shown in Fig. 1.

3 Results

In order to evaluate the accuracy of the proposed method; we follow the standard performance assessment procedures utilized in the video quality expert's group (VQEG) FR-TV Phase II test [5]. The objective and subjective scores [5], are fitted with the logistic function. Five parameters non-linear mapping (θ_1, θ_2, θ_3, θ_4 and θ_5) are utilized to change the set of quality ratings by the objective quality measures to a set of the predicted Difference Mean Opinion Score (*DMOS/MOS*) values denoted $DMOS_P/MOS_P$.

In equation (9), the logistic regression function is introduced which is employed for the nonlinear regression.

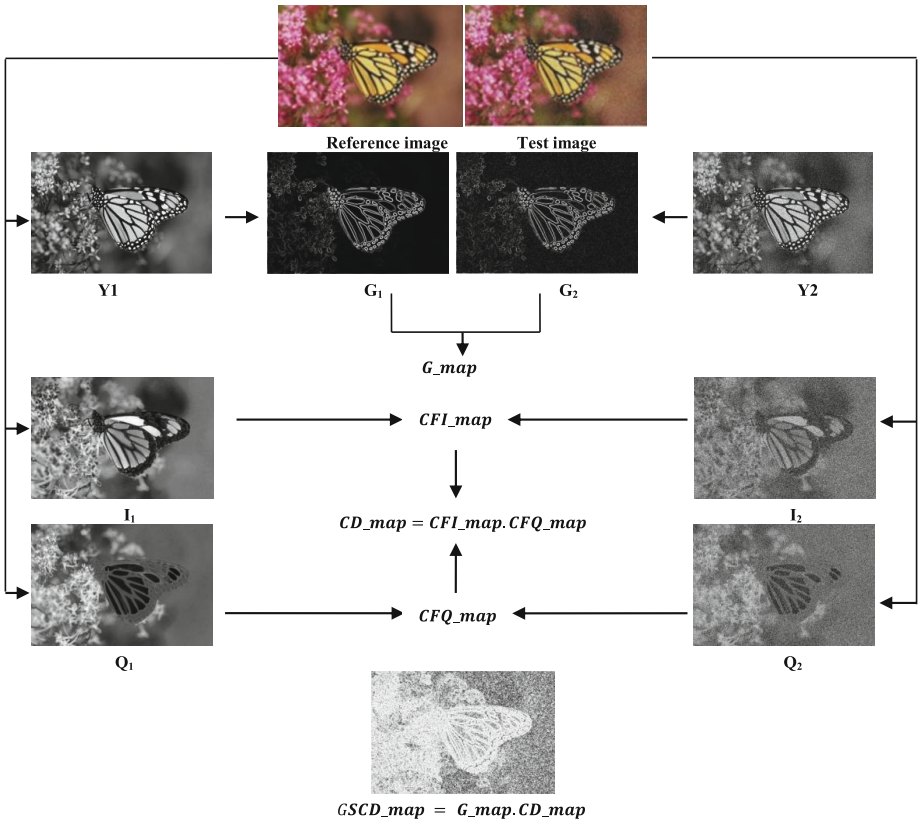


Fig. 1. Image quality assessment method

$$f(VQR) = \theta_1\left(\frac{1}{2} - \frac{1}{\exp(\theta_2(VQR - \theta_3))}\right) + \theta_4 VQR + \theta_5 \tag{9}$$

Where *VQR* is the value of the objective method and θ_1, θ_2, θ_3, θ_4, θ_5 are selected for the most excellent fit.

In this test, four metrics are used [26]: the Root mean square prediction error (*RMSE*), the Spearman rank-order correlations coefficient (*ROCC*), Kendall rank-order correlation coefficient (*KROCC*) and The Pearson linear correlation coefficient (*CC*). *ROCC* and *KROCC* evaluate the prediction monotonicity. *CC* and *RMSE* assess the prediction accuracy. *ROCC, KROCC* and *CC* are better with values closer to 1 or -1. Thus, *RMSE* is better when its values are small.

The first index *CC* (Pearson linear correlation coefficient) is defined by:

$$CC = \frac{\sum_{i=1}^{n}(DMOS(i) - \overline{DMOS})(DMOS_p(i) - \overline{DMOS_p})}{\sqrt{\sum(DMOS(i) - \overline{DMOS})^2}\sqrt{\sum(DMOS_p(i) - \overline{DMOS_p})^2}} \tag{10}$$

Where the index *i* denotes the image sample and *n* denotes the number of samples.

The second index is the Spearman rank-order correlations coefficient (*ROCC*); it is defined by:

$$ROCC = 1 - \frac{6\ \sum(DMOS(i) - DMOS_p(i))^2}{n(n^2 - 1)} \tag{11}$$

The third index is Kendall rank-order correlation coefficient (*KROCC*) [25]. It is designed to capture the association between two ordinal variables. Its estimate can be expressed as follows:

$$KROCC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} sgn(DMOS(i) - DMOS(j))sgn(DMOS_p(i) - DMOS_p(j))}{n(n-1)} \tag{12}$$

where:

$$sgn\big(DMOS(i) - DMOS(j)\big) = \begin{cases} 1\ if\ \big(DMOS(i) - DMOS(j)\big) > 0 \\ 0\ if\ \big(DMOS(i) - DMOS(j)\big) = 0 \\ -1 if\ \big(DMOS(i) - DMOS(j)\big) < 0 \end{cases}$$

and

$$sgn\Big(DMOS_p(i) - DMOS_p(j)\Big) = \begin{cases} 1\ if\ \Big(DMOS_p(i) - DMOS_p(j)\Big) > 0 \\ 0\ if\ \Big(DDMOS_p(i) - DMOS_p(j)\Big) = 0 \\ -1 if\ \Big(DMOS_p(i) - DMOS_p(j)\Big) < 0 \end{cases}$$

The forth one is the Root mean square prediction error (*RMSE*) between subjective (*DMOS*) and objective ($DMOS_P$) scores. It is defined by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(DMOS(i) - DMOS_p(i)\big)^2} \tag{13}$$

To judge the performance of the proposed approach, four kinds of databases are used: TID2008 database [9], CSIQ database [10], LIVE database [11] and TID2013 database [12]. The characteristics of these four databases are summarized in table 3.

The performance of GSCD metric is compared with PSNR, SSIM [2,16], Multis-cale-SSIM (MS-SSIM) [18,16], Visual Singal-to-Noise Ratio (VSNR) [19,16], Visual Information Fidelity (VIF) [20,16], Information Fidelity Criterion (IFC) [21,16], Noise Quality Measure (NQM) [22, 16], DTex [6], GSM [17], MAD [23,24], ERDDM [4], GSMD [14] and FSIM [13].

A comparative study of Sobel, Perwitt, Scharr and proposed operator is presented in Table 1 (TID2008 database is used in this experience), from which proposed operator could accomplish better performance than the other three. Furthermore, the choice

of *YIQ* color space needs to be proved. To this end, we run the proposed method with different four color spaces. The results are summarized in table 2 (TID2008 database is used in this experience).

Table 1. ROCC and KROCC values using four gradient operators

Gradient operator	Sobel	Perwitt	Scharr	Proposed operator
ROCC	0.8983	0.8996	0.8963	**0.9000**
KROCC	0.7143	0.7171	0.7104	**0.7175**

Table 2. ROCC and KROCC values using four color spaces

Color space	Lab	ycbcr	HSV	YIQ
ROCC	0.7684	0.8937	0.2983	**0.9000**
KROCC	0.5789	0.7110	0.2125	**0.7175**

The classification of the performance of all measures according to their ROCC values is presented in Table 8 reveal the reliability of the GSCD. Tables 4, 5, 6 and 7 show the obtained results. The top three measures for each assessment measure are highlighted in bold. We can see that the top methods are mostly GSCD, GMSD, FSIM and MAD. GSCD correlates much better with the subjective results than the other measures. Looking at the curves (Fig.2), the GSCD values are very close to DMOS and MOS, proving the efficiency of this measure.

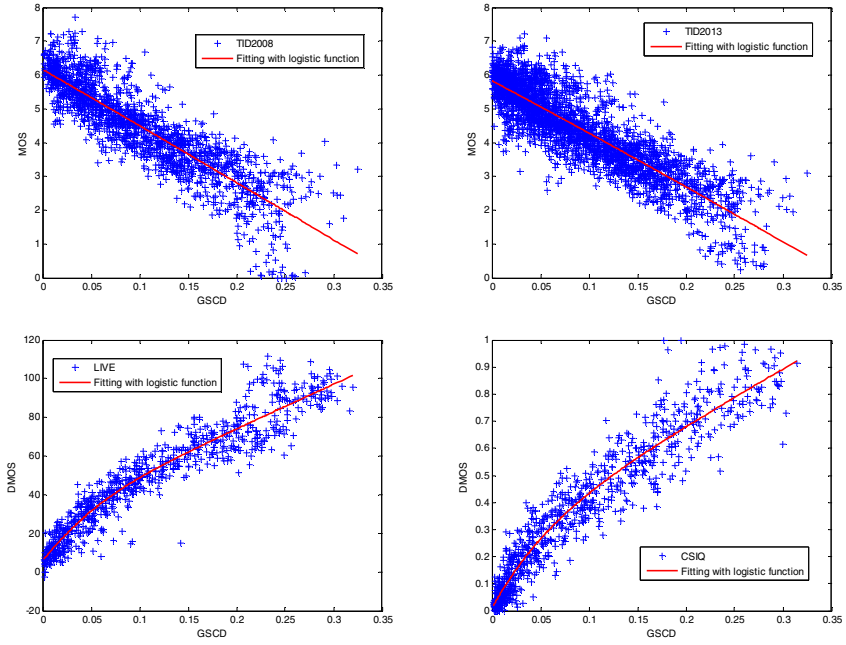


Fig. 2. Scatter plots of subjective scores versus scores from the proposed scheme on IQA databases

Moreover, an interesting result is obtained from the comparison of the GSCD with GMSD, FSIM and MAD in Tables 5 (TID2008 database). The values of ROOC are close to 1; this means that GSCD has a similar performance as the methods or earlier works. Results clearly indicate that our GSCD measure performs quite well and is competitive with other IQA measures.

In addition, to compare the efficiency of different models, the average execution time required an image of size 512×384 is calculated (the image is taken from TID2008 database). All metrics were run on a TOSHIBA Satetillete T130-11U notebook with Intel Core U4100 CPU@1.30 GHz and 3G RAM. The software platform used to run all metrics was MATLAB R2007a (7.4.0). Table 8 shows the required time in seconds per image. It is shown in Table 9 that the proposed measure takes more time than the PSNR, the GMSD, and the SSIM and it is faster than the Fsim.

VIF, VSNR, IFC, MS-SSIM, GSM, MAD, DCTex, NQM and ERDDM also take much longer processing time than the proposed method.

Moreover, we adjusted the parameters based on a dataset of TID2008 database. The adjusting measure was that the parameters values giving to a higher ROCC would be chosen. As a result, the parameters required in the proposed method were set as: C_1= 100, C_2=2050.

Table 3. Four databases and their characteristics

Database	Source Images	Distorted Images	Distortion Types	Image Type	Observers
TID2008	25	1700	17	color	838
CSIQ	30	866	6	color	35
LIVE	29	779	5	color	161
TID2013	25	3000	25	color	971

Table 4. Performance comparison for image quality assessment measures on live database

Method	*ROCC*	*KROCC*	*CC*	*RMSE*
PSNR	0.8756	0.6865	0.8723	13.3597
SSIM	0.9479	0.7963	0.9449	8.9454
MS-SSIM	0.9513	0.8044	0.9409	9.2593
VSNR	0.9280	0.7625	0.9237	10.4694
VIF	**0.9632**	**0.8270**	**0.9598**	**7.6670**
IFC	0.9259	0.7579	0.9268	10.2643
NQM	0.9086	0.7413	0.9122	11.1926
ERDDM	0.9496	0.8128	0.9619	6.3204
DCTex	0.9483	0.8066	0.9443	8.9897
GSM	0.9554	0.8131	0.9437	9.0376
MAD	**0.9669**	**0.8421**	**0.9674**	**6.9235**
Fsim	**0.9645**	**0.8363**	**0.9613**	**7.5296**
GMSD	0.9603	0.8271	0.9603	7.622
GSCD	0.9596	0.8222	0.9538	8.2074

Table 5. Performance comparison for image quality assessment measures on TID2008 database

Method	*ROCC*	*KROCC*	*CC*	*RMSE*
PSNR	0.5794	0.4210	0.5726	1.1003
SSIM	0.7749	0.5768	0.7710	0.8546
MS-SSIM	0.8542	0.6568	0.8451	0.7173
VSNR	0.7049	0.5345	0.6823	0.9810
VIF	0.7496	0.5868	0.8090	0.7888
IFC	0.5675	0.4236	0.7340	0.9113
NQM	0.6243	0.4608	0.6142	1.0590
ERDDM	0.5961	0.4411	0.6685	0.998
DCTex	0.4973	0.4095	0.5605	1.1113
GSM	0.8554	0.6651	0.8462	0.7151
MAD	0.8340	0.6445	0.8306	0.7474
Fsim	**0.8840**	**0.6991**	**0.8762**	**0.6468**
GMSD	**0.8907**	**0.7094**	**0.8788**	**0.6404**
GSCD	**0.9000**	**0.7175**	**0.8830**	**0.629**

Table 6. Performance comparison for image quality assessment measures on TID2013 database

Method	*ROCC*	*KROCC*	*CC*	*RMSE*
PSNR	0.6396	0.4698	0.669	0.9214
SSIM	0.7417	0.5588	0.7895	0.7608
MS-SSIM	0.7859	0.6047	0.8329	0.6861
VSNR	0.6812	0.5084	0.7402	0.8392
VIF	0.6769	0.5147	0.7720	0.7880
IFC	0.5389	0.3939	0.5538	1.0322
NQM	0.6432	0.474	0.6858	0.9023
ERDDM	0.5623	0.4124	0.6352	1.230
DCTex	0.5863	0.4573	0.6495	0.9425
GSM	0.7946	0.6255	0.8464	0.6603
MAD	0.7807	0.6035	0.8267	0.6975
Fsim	**0.8510**	**0.6665**	**0.8769**	**0.5959**
GMSD	**0.8044**	**0.6343**	**0.859**	**0.6346**
GSCD	**0.8681**	**0.6855**	**0.8819**	**0.5844**

Table 7. Performance comparison for image quality assessment measures on CSIQ database

Method	*ROCC*	*KROCC*	*CC*	*RMSE*
PSNR	0.8005	0.5984	0.7998	0.1576
SSIM	0.8756	0.6907	0.8612	0.1334
MS-SSIM	0.9133	0.7393	0.8990	0.1150
VSNR	0.8104	0.6237	0.7993	0.1578
VIF	0.9195	0.7537	0.9277	0.0980
IFC	0.7671	0.5897	0.8384	0.1431
NQM	0.7402	0.5638	0.7433	0.1756

Table 7. (*Continued*)

ERDDM	0.8626	0.6781	0.8295	0.1466
DCTex	0.8042	0.6420	0.7915	0.1605
GSM	0.9126	0.7403	0.8979	0.1156
MAD	**0.9467**	**0.7970**	**0.9502**	**0.0818**
Fsim	0.9310	0.7690	0.9192	0.1034
GMSD	**0.957**	**0.8133**	**0.9541**	**0.0786**
GSCD	**0.9602**	**0.8194**	**0.9578**	**0.0755**

Table 8. Ranking of IQA metrics' performance on four databases

Method	Live	TID2008	TID2013	CSIQ
PSNR	14	12	11	12
SSIM	10	7	7	8
MS-SSIM	7	5	5	6
VSNR	11	9	8	10
VIF	**2**	8	9	5
IFC	12	13	14	13
NQM	13	10	10	14
ERDDM	8	11	13	9
DCTex	9	14	12	11
GSM	6	4	4	7
MAD	**1**	6	6	**3**
Fsim	**3**	**3**	**2**	4
GMSD	4	**2**	**3**	**2**
GSCD	**5**	**1**	**1**	**1**

Table 9. Running time of the competing IQA models

Method	Time (second)	Method	Time (second)
PSNR	0.0493	ERDDM	9.6089
SSIM	0.1917	DCTex	0.5327
MS-SSIM	1.1304	GSM	1.4003
VSNR	1.5018	MAD	15.6235
VIF	5.1429	Fsim	2.4990
IFC	4.6738	GMSD	0.1602
NQM	1.8846	GSCD	0.4361

4 Conclusion

This paper describes an efficient method for image quality assessment. Its main feature is that this new method uses the gradient similarity and color distorted measure. The reference and test images are transformed respectively using color distorted and

gradient mask. The difference between the reference and test images is computed using simple function. A comparative study has been carried in this work.

The obtained results are competitive with the previous works.

Future works following this study will include the use of others characteristics to assess image quality.

References

1. Yang, C., Kwok, S.H.: Efficient gamut clipping for color image processing using LHS and YIQ. Opt. Eng. 42(3), 701–711 (2003)
2. Wang, Z., Bovik, A.C., Sheikh, H.R., Simocelli, E.P.: Image quality assessment: From error measurement to structural similarity. IEEE Trans. Image Processing 13(4), 600–612 (2004)
3. Guan-Hao, C., Chun-Ling, Y., Sheng-Li, X.: Gradient-based structural similarity for image quality assessment. In: Proc. ICIP 2006, pp. 2929–2932 (2006)
4. Ahmed Seghir, Z., Hachouf, F.: Edge-region information measure based on deformed and displaced pixel for Image Quality Assessment. Signal Processing: Image Communication 26(8-9), 534–549 (2011)
5. Final VQEG report on the validation of objective quality metrics for video quality assessment: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/
6. Zhang, F., Ma, L., Li, S.: Practical image quality metric applied to image coding. IEEE Trans. Multimedia 13, 615–624 (2011)
7. Jain, R., Kasturi, R., Schunck, B.G.: Machine Vision. McGraw-Hill, NewYork (1995)
8. Jähne, B., Haubecker, H., Geibler, P.: Handbook of Computer Vision and Applications. Academic, New York (1999)
9. Ponomarenko, N., Egiazarian, K.: Tampere Image Database, TID 2008, http://www.ponomarenko.info/tid2008.htm
10. Larson, C., Chandler, D.M.: Categorical Image Quality (CSIQ) Database 2009, http://vision.okstate.edu/csiq
11. Sheikh, H.R., Seshadrinathan, K., Moorthy, A.K., Wang, Z., Bovik, A.C., Cormack, L.K.: Image and Video Quality Assessment Research at LIVE 2004 (2004), http://live.ece.utexas.edu/research/quality
12. Ponomarenko, N., et al.: Color image database TID2013: Peculiarities and preliminary results. In: Proc. 4th Eur. Workshop Vis. Inf. Process., pp. 106–111 (June 2013)
13. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing 20(8), 1–26 (2011)
14. Xue, W., Zhang, L., Mou, X., Bovik, A.C.: Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. Presented at IEEE Transactions on Image Processing, 684–695 (2014)
15. Kovesi, P.: Image features from phase congruency. Videre: Journal of Computer Vision Research 1(3), 1–26 (1999)
16. Gaubatz, M.: Metrix MUX Visual Quality Assessment Package: MSE, PSNR, SSIM, MSSIM, VSNR, VIF, VIFP, UQI, IFC, NQM, WSNR, SNR
17. http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
18. Liu, A., Lin, W., Narwaria, M.: Image quality assessment based on gradient similarity. IEEE Transactions on Image Processing 21(4), 1500–1512 (2012)

19. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: Proc. IEEE Asilomar Conf. Signals, Syst., Comput., Pacific Grove, CA, pp. 1398–1402 (November 2003)
20. Chandler, D.M., Hemami, S.S.: VSNR: A wavelet-based visual signal-to-noise-ratio for natural images. IEEE Trans. Image Process. 16(9), 2284–2298 (2007)
21. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. 15(2), 430–444 (2006)
22. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. on Image Processing 14(12), 2117–2128 (2005)
23. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on degradation model. IEEE Trans. on Image Processing 9(4), 636–650 (2000)
24. Larson, E.C., Chandler, D.M.: Most apparent distortion: Full-reference image quality assessment and the role of strategy. J. Electron. Imaging 19(1), 011006:1–011006:21 (2010)
25. Larson, E., Chandler, D.: Full-Reference Image Quality Assessment and the Role of Strategy: The Most Apparent Distortion, http://vision.okstate.edu/mad/
26. Chok, N.S.: Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data. Master's Thesis, University of Pittsburgh (2010)
27. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. 20(5), 1185–1198 (2011)

Computational Intelligence: Machine Learning

Biomarker Discovery Based on Large-Scale Feature Selection and MapReduce

Ahlam Kourid(✉) and Mohamed Batouche

Computer Science Department, College of NTIC, Constantine 2 University – A. Mehri, 25000, Constantine, Algeria
ahlem.kou@gmail.com, mohamed.batouche@univ-constantine2.dz

Abstract. Large-scale feature selection is one of the most important fields in the big data domain that can solve real data problems, such as bioinformatics, where it is necessary to process huge amount of data. The efficiency of existing feature selection algorithms significantly downgrades, if not totally inapplicable, when data size exceeds hundreds of gigabytes, because most feature selection algorithms are designed for centralized computing architecture. For that, distributed computing techniques, such as MapReduce can be applied to handle very large data. Our approach is to scale the existing method for feature selection, Kmeans clustering and Signal to Noise Ratio (SNR) combined with optimization technique as Binary Particle Swarm Optimization (BPSO). The proposed method is divided into two stages. In the first stage, we have used parallel Kmeans on MapReduce for clustering features, and then we have applied iterative MapReduce that implement parallel SNR ranking for each cluster. After, we have selected the top ranked feature from each cluster. The top scored features from each cluster are gathered and a new feature subset is generated. In the second stage, the new feature subset is used as input to the proposed BPSO based on MapReduce which provides an optimized feature subset. The proposed method is implemented in a distributed environment, and its efficiency is illustrated through analyzing practical problems such as biomarker discovery**.**

Keywords: Feature selection · Large-scale machine learning · Big data analytics · Bioinformatics · Biomarker discovery

1 Introduction

With the progress of high technology in several fields that produce an important volume of data such as Microarray and Next generation sequencing in bioinformatics [1], deal with high dimensional data becomes a challenge for several tasks in machine learning. Feature selection is one of the techniques of reduction dimensionality [2] that is effective in removing irrelevant data; increasing learning accuracy, therefore becomes very necessary for machine learning tasks. Scalability can become a problem for even simple and centralized approaches, for that feature selection methods based on parallel algorithm will be the mainly choice for dealing with large-scale data. Many parallel algorithms are implemented using different parallelization techniques

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 81–92, 2015.
DOI: 10.1007/978-3-319-19578-0_7

such as MPI (The Message Passing Interface), and MapReduce. MapReduce is a programming model for distributed computation, derived from the functional programming concepts, and is proposed by Google for large-scale data processing in a distributed computing environment [3].

Recent comparisons studies of feature selection methods in high-dimensional data have shown that the combination of K-means clustering and filter method based SNR (Signal to Noise Ratio) score combined with binary PSO is a graceful method for classification problem [4]. The method is applied for classification of DNA microarray data. To resolve redundancy in gene expression values one approach i.e. sample based clustering by using k-means clustering algorithm is used and the genes (features) are being grouped into number of clusters. After clustering SNR ranking is being used to rank each gene (feature) in every cluster. The gene subset selected by taking the top scored gene (feature) from each cluster is validated with an SVM classifier, and will be taken as the initial search space to find the optimized subset by applying PSO and the optimized subset is used to train different classifier such as SVM [4]. However, the existing method is limited over large scale datasets. In order to overcome that problem we present our method that is suitable for very large data and that has the potential for parallel implementation, based on parallel Kmeans on MapReduce for clustering a huge amount of features, so similar features having the same characteristics will be grouped in the same cluster, and on an iterative MapReduce that implement parallel SNR ranking for each cluster. Finally, the top non-redundant ranked features selected are input to BPSO on MapReduce to select the relevant features.

2 Parallel Programming Paradigm and Framework

In order to implement our approach to cope with large scale data sets, we are using Hadoop platform and MapReduce as parallel programming paradigm .

2.1 MAPREDUCE

MapReduce is a functional programming model that is well suited to parallel computation. The model is divided into two functions which are map and reduce .In MapReduce; all data are in the form of keys with associated values. The following notation and example are based on the original presentation [3]:

A. **Map Function**

A map function is defined as a function that takes a single key-value pair and outputs a list of new key-value pairs. The input key may be of a different type than the output keys, and the input value may be of a different type than the output values:

$$\text{Map} :(\text{K1}, \text{V1}) \rightarrow \text{list} ((\text{K2}, \text{V2})) \quad (1)$$

B. **Reduce Function**

A reduce function is a function that reads a key and a corresponding list of values and outputs a new list of values for that key. The input and output values are of the same type.

$$\text{Reduce :(K2, list (V2))} \rightarrow \text{list (V2)} \tag{2}$$

2.2 HADOOP Platform

Hadoop is an open source Java based framework to store and process large amounts of data. It allows distributed processing of data which is present over clusters using functional programming model. MapReduce is the most important algorithm implemented in Hadoop. Each Map and Reduce is independent of other Maps and Reduces. Processing of data is executed in parallel to other processes. A job scheduler or job tracker tracks MapReduce jobs which are being executed. Tasks like Map, Reduce and Shuffle are accepted from Job Tracker by a node called Task Tracker. Hadoop architecture is defined as follows: Hadoop consists of two components, the Hadoop Distributed File System (HDFS) and MapReduce, performing distributed processing by single-master and multiple-slave servers. There are two elements of MapReduce, namely JobTracker and TaskTracker, and two elements of HDFS, namely DataNode and NameNode. [5].

3 Scaling Up Feature Selection Algorithm

For scaling up the existing method for feature selection, we propose an approach based on MapReduce which is composed of two stages. The first stage consists in filtering the set of features by selecting the top scored features whereas the second stage optimizes the obtained subset of selected features.

3.1 Filtering the Set of Features

This stage is scalable and implements K-means clustering on MapReduce and SNR ranking on MapReduce for each cluster. It is designed for the purpose of eliminating redundancy in features and selecting the top scored features [6]. And it is composed of the following steps:

Step1: clustering features (genes) with parallel K-means on MapReduce. As by applying clustering technique we can group similar type of features in the same cluster, so that best features from each cluster can be selected.
Step2: mappers read lines (features) and compute SNR score for each feature.
Step3: according to the paradigm shuffle and sort in MapReduce, the final output file contains ranked SNR values. Top ranked features are selected in two cases:

- One output file: the top ranked features are selected from this file.
- Multiple output files: each file is ranked by SNR value, for that Terasort can be used to rank all SNR values from these files. Terasort is a standard map/reduce sort, and it is implemented as benchmark in hadoop [7].

Step4: After that the best scored feature in a cluster is selected, and go to step 2 for the next cluster. We can assure that applying SNR and selecting the best scored feature from each cluster the resultant feature gene subset have no redundancy.
Step5: top features (genes) ranked from each cluster are aggregated and validated with SVM classifier using the evaluation method 10 foldCV.

The system architecture for the proposed method in stage-I- is illustrated in **Fig.1**.

3.2 Optimizing the Subset of Selected Features

This stage aims to select an optimized subset of features from the subset selected in the previous stage. It is parallel, and can provide scalability to a certain degree because of the SVM classifier which is sequential. In this stage; we have used four MapReduce jobs described by the following steps:

Step1: the subset of features selected and validated in the previous stage, is the input to the novel BPSO proposed based on MapReduce, we have divided the particles of swarm into groups, so that the input file contains particles defined by their groups, in the first MapReduce job, mappers evaluate fitness (accuracy of SVM) of particles in parallel.
Step2: in the second MapReduce job, mappers read output file from the first job and emit the group identifier as key in order to group particles. Reducers evaluate Gbestg of each group in parallel, and emit "one" as key and the Gbestg of the group with fitness of Gbestg of the group as value.
Step3: the third MapReduce job evaluates the Gbestglobal, which is the maximum of all Gbestg of each Group. The output file of this job contains the Gbestglobal and its fitness.
Step4: the file output of the first job in HDFS is the input of the fourth job, in this job mappers read the output file of the third job that contains Gbestglobal and its fitness from HDFS, in order to evaluate the new positions and the new velocities in parallel. The output of this job is the new swarm for the next iteration.

The system architecture for the proposed method in stage-II- is illustrated in **Fig.2**.

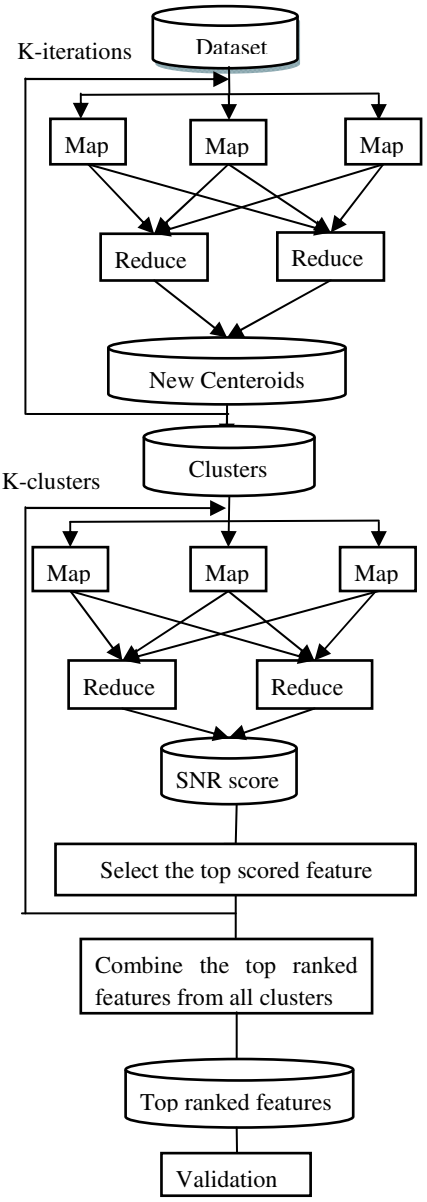


Fig. 1. System architecture of the proposed method stage-I-

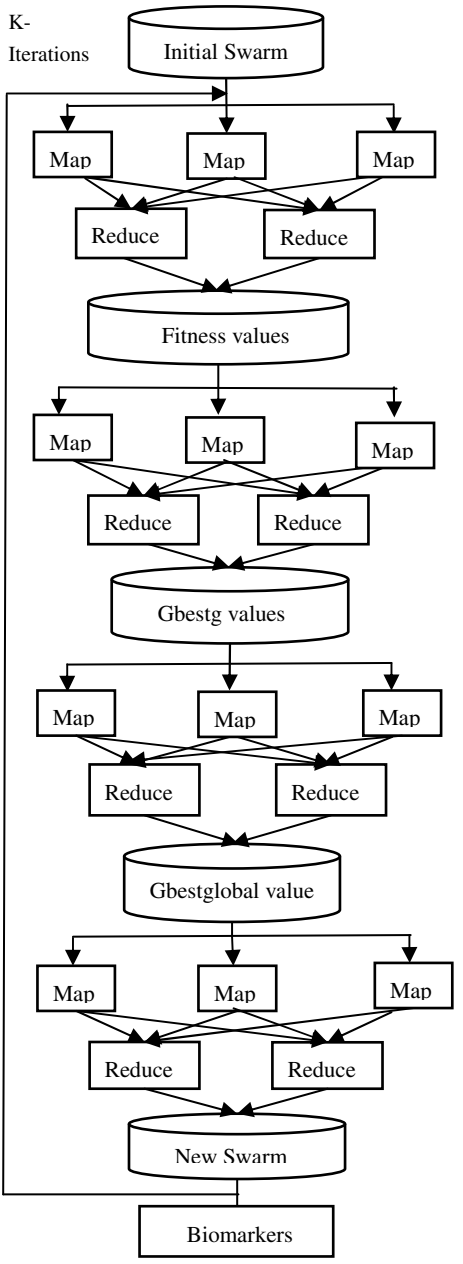


Fig. 2. System architecture of the proposed method stage-II-

4 Implementation of the Proposed Approach

In order to implement the proposed approach for scaling up the existing method for feature selection, we describe in the following the algorithms and map/reduce functions.

4.1 The First Stage

In this stage, we are using K-means along with SNR ranking on MapReduce which are defined as follows:

Algorithm 1. Kmeans on MapReduce.

Input : Training data (features) .
Output: Clusters.

Algorithm 1.1. k-means::Map

Input: Training data x ∈ D, number of clusters k, distance measure d
1: If first Map iteration **then**
2: Initialize the k cluster centroids C randomly
3: Else
4: Get the k cluster centroids C from the previous Reduce step.
5: Set $S_j = 0$ and $n_j = 0$ for j = {1, ... , k}
6: For each $\mathbf{x_i} \in$ **D do**
7: y_i = arg $\min_j d(\mathbf{x_i}, \mathbf{c_j})$
8: $S_{y_i} = S_{y_i} + x_i$
9: $n_{y_i} = n_{y_i} + 1$
10:For each j ∈ {1, … , k} **do**
11: Output(j, <S_j, n_j>)

Algorithm 1.2. k-means::Reduce

Input : List of centroid statistics – partial sums and counts [<S_j^l , n_j^l>] – for each centroid j ∈ {1, ..., k}
1:For each j ∈ {1, ..., k} **do**
2: Let λ be the length of the list of centroid statistics
3 : $n_j = 0$, $S_j = 0$
4 : **For each** l ∈ {1, ..., k} **do**
5 : $n_j = n_j + n_j^l$

6 : $S_j = S_j + S_j^l$

7 : $\mathbf{c_j} = \frac{S_j}{n_j}$

8 : Output (j, $\mathbf{c_j}$)

The whole clustering is run by a Master, which is responsible for running the Map (cluster assignment) and Reduce (centroid re-estimation) steps iteratively until k-means converges [8].

Algorithm 2. SNR on MapReduce

Input : Clusters .
Output: Feature subset of top scored features from clusters.

- List: contains target classes of samples in order.
- Record: contains values of samples for $feature_i$.
- DFS: is a distributed directory system for storage of output and input files of MapReduce.
- ID_feature: is an identifier characterizes each feature.
- $file\_cluster_i$:contains features of cluster i.

Clustering features withAlgorithm 1.
For each cluster i do
DFS.put ($file\_cluster_i$)

Map function (parallel over features) (key: ID_feature, value: record)
List= [class1, class2, class2……]
Iterate over record and list
compute μ1, μ2
compute σ1, σ2
compute SNR
Output (SNR, (ID_feature, record))

Reduce Function (key: SNR, value :(ID_feature, record)
Output (SNR, (ID_feature, record))
Select top scored feature.
DFS.delete ($file\_cluster_i$).

Aggregation and validation of the top scored features selected.

4.2 The Second Stage:

In this stage, we are using Binary PSO on MapReduce which is composed of four MapReduce jobs. In Hadoop, a mechanism of JobControl classes is provided to execute the four jobs sequentially.

Algorithm 3. PSO on MapReduce

Input : Initial swarm of particles and the subset of top features selected and validated .
Output: Best solution Gbest.

- GroupP: we have defined at the beginning several groups of particles, GroupP is the identifier of each group.
- P: position of a particle, Pbest: best position of a particle, FitPbest: fitness of Pbest, Gbest: global position of particles, FitGbest: fitness of Gbest, V: velocity of a particle.

First job
Mapper (key: ID_particle, value: (P, Pbest, FitPbest, Gbest, FitGbest, V, GroupP) (parallel mappers)
Initial Pbest, FitPbest, Gbest, FitGbest are empty.
fitness (): function of evaluation of the fitness of the designed particle (accuracy SVM) and take as input P and features selected.
If fitness (P) >FitPbest
Pbest=P.
FitPbest= fitness (P).
Emit(ID_particle, (P, Pbest, FitPbest, Gbest, FitGbest, V, GroupP))
Reducer (key: ID_particle, value: (P, Pbest, FitPbest, Gbest, FitGbest, V,GroupP) (parallel reducers)
Emit(ID_particle, (P, Pbest, FitPbest, Gbest, FitGbest, V,GroupP)) (file-output1 in HDFS)

Second job
Mapper (key: ID_particle, value: (P, Pbest, FitPbest, Gbest, FitGbest, V, GroupP) (parallel mappers)
Emit(GroupP, (ID_particle ,P, Pbest, FitPbest, Gbest, FitGbest, V))
Reducer (key: GroupP, value: (ID_particle, P, Pbest, Gbest, FitGbest, V) (parallel reducers)
Initial Gbestg is empty.
Cpt: number of 1 in Gbest, initialized to 0.
For all values
Gbestg = maximum of all Gbest with minimum number of Cpt (in case of equality between Gbest).
FitGbestg= FitGbest of Gbestg.
Emit (ONE, (Gbestg, FitGbestg)) (file-output2 in HDFS)

Third job
Mapper (key: ONE, value: (Gbestg, FitGbestg) (parallel reducers)
Emit (ONE, (Gbestg, FitGbestg))
Reducer (key: ONE, value: (Gbestg, FitGbest) (parallel reducers)
Initial Gbestglobal is empty. Cpt1: number of 1 in Gbestg, initialized to 0.
For all values
Gbestglobal = maximum of Gbestg with minimum number of Cpt1 (in case o equality between Gbestg)
FitGbestglobal= FitGbest of Gbestglobal.
Emit (Gbestglobal, (FitGbestglobal)) (file-output3 in HDFS)

Fourth job
Mapper (key: ID_particle, value: (P, Pbest, FitPbest, Gbest, FitGbest, V, GroupP) (parallel mappers)
Read file-output3 from HDFS
Gbest = Gbestglobal
FitGbest = FitGbestglobal
V'= New_Velocity (V, P, Pbest, Gbest) /* New_Velocity is a function for the evaluation of the new velocity*/
P'= New_Position (P, V') /* New_Position is a function for the evaluation of the new position*/
P=P', V=V'
Emit(ID_particle, (P, Pbest, FitPbest, Gbest, FitGbest V, GroupP))
reducer (key: ID_particle, value: (P, Pbest, FitPbest, Gbest, FitGbest, V, GroupP) (parallel reducers)
Emit(ID_particle, (P, Pbest, FitPbest, Gbest, FitGbest, V, GroupP))

Repeat the execution of jobs K-iterations.

5 Results and Experiments

We have used tow datasets of cancer RNA-seq gene expression data (gastric cancer, ESCA (esophageal carcinoma)): gastric dataset derived from the main source of gene expression data Omnibus. The last, ESCA derived from TCGA (Cancer Genome Atlas), and four gene expression microarray datasets (tow ovarian cancer datasets, gastric cancer dataset, ESCC dataset (esophageal squamous cell carcinoma)) derived from Omnibus. Our approach is implemented on two-node cluster (master and slave), both master machine and slave machine are equipped with dual core processor and 4GB RAM memory for master node, and 2 GB for slave node. The operating system installed on the two nodes is Linux Ubuntu 13.10. The experiment is done using hadoop-1.2.1 and mahout 0.9 [9]. The cluster is configured in fully-distributed mode [10]. We have used support vector machine (SVM) to obtain classification accuracy, and the cross validation method 10 foldCV for performance evaluation of the classifier SVM. In order to improve the scalability of our method we have used a synthetic dataset (duplicate genes of each dataset), the size of data increased reaches 5GB for each dataset. Experiment is done with 5 clusters and 10 clusters. The performance of

our method is compared to other approaches in the literature: an approach Based on Neighborhood Rough Set and Probabilistic Neural Networks Ensemble is proposed for the classification of Gene Expression Profiles [11], in [12] authors proposed a new selection method of interdependent genes via dynamic relevance analysis for cancer diagnosis. However, in the work presented in [13] a sequential forward feature selection algorithm to design decision tree models is suggested for the identification of biomarkers for Esophageal Squamous Cell Carcinoma. The obtained results are shown on Table 1, Table 2 and Table 3.

Table 1. Accuracy of SVM and number of genes selected in our method with normal datasets and comparison with other approaches

dataset	Ng	BPSO on MapReduce				[11]		[12]		[13]	
		Se	**Sp**	**Acc**	**#**	**Acc**	**#**	**Acc**	**#**	**Acc**	**#**
Ovrian [11]	15154	1	0,98	99	3	96	9	-	-	-	-
Gastric [12]	4522	1	1	100	2	-	-	96	14	-	-
ESCC [13]	22477	0,96	0,96	96	2	-	-	-	-	97	2
Ovrian	54675	1	1	100	2	/		/		/	
Gastric	21475	1	1	100	1						
ESCA	26540	1	1	100	2						

Ng: number of genes, **Se:** sensibility, **Sp:** specificity, **Acc:** accuracy (%), **#:** number of genes selected.

Table 2. Accuracy of SVM and number of genes selected in our method with large-scale datasets and comparison with other approaches

dataset	Size dataset	BPSO on MapReduce				[11]		[12]		[13]	
		Se	**Sp**	**Acc**	**#**	**Acc**	**#**	**Acc**	**#**	**Acc**	**#**
Ovarian [11]	5GB	1	0,98	99	3	96	9	-	-	-	-
Gastric [12]	5GB	1	1	100	2	-	-	96	14	-	-
ESCC [13]	5GB	0,96	0,96	96	2	-	-	-	-	97	2
Ovarian	5GB	1	1	100	2	/		/		/	
Gastric	5GB	1	1	100	1						
ESCA	5GB	1	1	100	2						

Table 3. List of biomarkers discovered

Type of cancer	Biomarkers	Related to cancer
Gastric cancer	**VSIG2** (V-set and immunoglobulin domain containing 2)	Selected from 22 gastric cancer biomarkers [14].
	D26129_at (RNS1 Ribonuclease A (pancreatic))	Considered among the non-regulated genes in gastric cancer [15].
	M62628_s_at (Alpha-1 Ig germline C-region membrane-coding region)	
Ovarian cancer	**METTL7A** (methyltransferaselike 7A)	Selected among the 28 genes markers linked to cancer [16].
	GALC (galactosylceramidase)	Selected Among the new differentially expressed genes in cell lines MKN45 gastric cancer [17].
Esophageal cancer	**ADAM12** (ADAM Metallopeptidase Domain 12)	Biomarkers of two types of cancer, breast cancer and bladder cancer [18].
	GPR155 (G protein-coupled receptor 155)	Melanomabiomarker for mouse [19].
	SH3BGRL (SH3 domain binding glutamate-rich protein)	Selected from 20 potential biomarkers of breast cancer [20]

6 Conclusion and Future Work

In this paper, we presented a large-scale feature selection based on MapReduce for biomarker discovery. From the obtained results and comparative analysis we can conclude that our method performs well, and gives better performance than centralized approaches. For that, our method can be applied to handle large-scale datasets and to overcome the challenge of feature selection in Big Data, especially for biomarker discovery in bioinformatics. Our method is auto-scalable and can be executed in a distributed environment with any number of nodes. Our future work is to implement our approach on Spark for better performance in time execution.

References

1. Jay, S., Hanlee, J.: Next-generation DNA sequencing. Nature Biotechnology 26(10), 1135–1145 (2008)
2. Yvan, S., Inaki, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
3. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI 2004: Sixth Symposium on Operating System Design and Implementation, Sponsored by USENIX, in Cooperation with ACM SIGOPS, pp. 137–150 (2004)
4. Barnali, S., Debahuti, M.: A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data. Procedia Engineering 38, 27–31 (2012)
5. Azli, A., et al.: Distributed visual enhancement on surveillance video with Hadoop Mapreduce and performance evaluation in pseudo distributed mode. Australian Journal of Basic and Applied Sciences 8(9), 38 (2014)
6. Kourid, A.: Iterative MapReduce for Feature Selection. International Journal of Engineering Research & Technology 3(7) (2014)
7. White, T.: Hadoop the definitive guide. O'Reilly Media (2012)
8. Bekkerman, R., Bilenko, M., Langford, J.: Scaling up Machine learning. Cambridge University Press (2011)
9. Sean, O., et al.: Mahout in action. Manning Publications (2011)
10. Gaizhen, Y.: The Application of MapReduce in the Cloud Computing. In: Intelligence Information Processing and Trusted Computing (IPTC), pp. 154–156. IEEE (2011)
11. Yun, J., Guocheng, X., Na, C., Shan, C.: A New Gene Expression Profiles Classifying Approach Based on Neighborhood Rough Set and Probabilistic Neural Networks Ensembl. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013, Part II. LNCS, vol. 8227, pp. 484–489. Springer, Heidelberg (2013)
12. Sun, X., et al.: Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis. Journal of Biomedical Informatics 46(2), 252–258 (2013)
13. Tung, C.W., et al.: Identification of Biomarkers for Esophageal Squamous Cell Carcinoma Using Feature Selection and Decision Tree Methods. The ScientificWorld Journal (2013)
14. Yang, S., Chung, H.C., et al.: Novel biomarker candidates for gastric cancer. Oncology Reports 19(3), 675–680 (2008)
15. Geetha Ramani, R., Gracia Jacob, S.: Benchmarking Classification Models for Cancer Prediction from Gene Expression Data: A Novel Approach and New Findings. Studies in Informatics and Control 22(2), 133–142 (2013)
16. Li, X., et al.: SSiCP: a new SVM based Recursive Feature Elimination Algorithm for Multiclass Cancer Classification. Bio-Medical Materials and Engineering 23, S1027–S1038 (2014)
17. Tuan, T.F., et al.: Putative tumor metastasis-associated genes in human gastric cancer. International Journal of Oncology 41(3), 1068–1084 (2012)
18. Fröhlich, C., et al.: Molecular Profiling of ADAM12 in Human Bladder Cancer. Clinical Cancer Research 12(24), 7359–7368 (2006)
19. Hacker, E., et al.: Reduced expression of IL-18 is a marker of ultraviolet radiation-induced melanomas. Int. J. Cancer 123(1), 227–231 (2008)
20. Mayer, M.: Breast Cancer Prognostic Biomarkers. Accelerating science (2014)

Social Validation of Solutions in the Context of Online Communities

An Expertise-Based Learning Approach

Lydia Nahla Driff(✉), Lamia Berkani, Ahmed Guessoum, and Abdellah Bendjahel

Artificial Intelligence Laboratory (LRIA), Department of Computer Science, USTHB, Bab Ezzouar, Algeria
driff.nahla@gmail.com, l_berkani@hotmail.com, {lberkani,aguessoum}@usthb.dz

Abstract. Online Communities are considered as a new organizational structure that allows individuals and groups of persons to collaborate and share their knowledge and experiences. These members need technological support in order to facilitate their learning activities (e.g. during a problem solving process).We address in this paper the problem of social validation, our aim being to support members of Online Communities of Learners to validate the proposed solutions. Our approach is based on the members' evaluations: we apply three machine learning techniques, namely a Genetic Algorithm, Artificial Neural Networks and the Naïve Bayes approach. The main objective is to determine a validity rating of a given solution. A preliminary experimentation of our approach within a Community of Learners whose main objective is to collaboratively learn the Java language shows that Neural Networks represent the most suitable approach in this context.

Keywords: Learning Community · Social Validation · Expertise-Based Learning · Machine Learning

1 Introduction

Today, with the great development of Information and Communication Technologies, a wide diversity of social learning frameworks have been promoted, including Online Learning Communities (OLCs) and social networks. The notion of OLC has been defined in different ways, exploring mainly the social aspects of collaborative learning (Laister and Kober, 2013) and the research and theory concerned with social support for learning (Swan and Shea, 2005).

One of the most important challenges of such communities is to enhance the knowledge exchange and sharing among the different members. With the increasing number of interactions, members collectively produce new knowledge in various formats (documents, solutions to problems, etc.) that will subsequently be published to the whole community (Le Boulch, 2009).The production of this knowledge is increasingly developed by members of the community who have different levels of expertise (experts,

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 93–104, 2015.
DOI: 10.1007/978-3-319-19578-0_8

novices, etc.) and this highlights the need for validation of the new knowledge before it is stored and published to the rest of the community so as to ensure its reliability.

Validation is the expression of a judgment on a concept or whatever needs to be assessed after study/observation. This judgment can be favorable or not. Social validation of a concept is a collective action that aims at the evaluation of this concept based on various judgments and opinions expressed by different people from the field (Herr and Anderson, 2008), on the basis of statistical analysis, or approved opinions, experiences, etc. Its main objective is the assessment whether the concept is good or not and this can be represented by a degree of validity which is the percentage of conformity of the concept.

We focus in this work on the social validation of the proposed solutions within an OLC. We aim to support members in this process, providing them with a tool that will help them to "automatically" validate newly proposed solutions. We address the need to ensure the credibility of the validation process and we propose an expertise-based learning approach, by reusing past experiences (i.e. previous validations made by supervisors).

The review of the literature about social validation shows that little work has focused on this aspect in an educational setting. We especially mention the work proposed by Cabana et al. (2010) about the social validation on collective annotations where the authors addressed the problem of scalability (i.e. a resource which is more and more annotated is less and less exploitable by individuals). The authors proposed a way to socially validate collective annotations with respect to the social theory of information. Berkani et al. (2013) proposed a social validation of learning objects based on two features: (1) the members' assessments, formalized semantically, and (2) an expertise-based learning approach, applying a machine learning technique. The authors used neural networks because of their proven efficiency in many domains such as complex problem solving.

The remainder of this paper is structured as follows. In Section 2, we present the context of the study. Our contribution is presented in Section 3, where we give a detailed description of the parameters that are related to the evaluation of a solution. Then, we present the application of the aforementioned Machine Learning techniques to any given solution using the defined parameters. The experimental results are presented in Section 4 where a discussion of the strengths and weaknesses of each technique is presented. The conclusion and perspectives are stated in Section 5.

2 Context of the Study

In our study, we consider a learning community related to the domain of higher education. The members of this community target the learning of the JAVA programming language and its different concepts. The main objective of this community is to improve the learners' skills and his acquisition of new knowledge.

We assume that the community includes members with different skill levels (beginners, advanced, experts, etc.) and the Java language includes various concepts (classes, abstraction, and so on), including the handling of tools (Java web GUI, JVM, etc.).

We have found out that the discussion forum and/or the Frequently Asked Questions are services that can be considered as the most used by the community members. Indeed, these often use these services in their interactions and information exchange.

However, one of the problems encountered by the community members is the validation of the proposed solutions; several questions can be asked at this point: Is a given solution correct? If yes, to what extent is it accurate? And, more importantly, who has validated it? Was it validated by a set of members, by a single expert, or within some other setting? This is why we focus in this work on the social validation of a solution, and we try to automate this process in order to support community members in their learning activities.

Automation is very cost-effective when it comes to time saving. It helps avoid several phases and replaces the manual work done by the supervisor (or teacher) whose role is to validate the solutions that are proposed by the community members (or learners). On the other hand, the social validation can give some precision about the obtained results by considering distinct opinions that are based on different criteria.

In the next section, we present our approach for automating the process of social validation, bearing in mind that we consider the validation of one solution at a time.

3 Contribution

In our work, we have tried to automate the validation process using Machine Learning (ML) techniques. The choice for these techniques has been dictated by the fact that they allow to take into account the rich experience (knowledge repository) in terms of validations of solutions throughout the lifetime of a community. As such, they directly take into account the existing experience, which makes them very different from conventional algorithmic methods where an exact and accurate understanding of the factors that are taken into account in any validation of a solution is required.

We start by the modelling of the solution validation problem in terms of some specific parameters that need to be represented. Our goal is to define an evaluation in a unique way so as to be able to manipulate it in the (automatic) training phase.

In this section, we present these parameters as well as the process followed by each of the considered ML techniques.

3.1 Parameters of an Evaluation

As explained above, the prediction of a degree of validity for a given solution is collective, based on the various members' assessments of the solution. To this end, we have defined the parameters that we consider as being the most important and significant ones to identify, characterize, and implement each assessment (see Fig. 1). Thus we believe that the evaluation by a given member M_i concerns the level of the evaluator, the assigned score, the confidence, the evaluation context, the skill level and the success. We now explain each of these parameters.

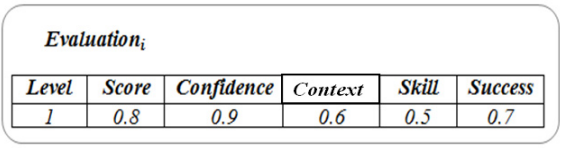


Fig. 1. Parameters of an evaluation

1. ***Level of the evaluator:*** Each assessment corresponds to a single member. This member has a certain amount of knowledge and expertise which are measured by the parameter "Level". In our case, a member is either a Professor, a PhD student, a Master's student, a student preparing a Bachelor's degree or a member with basic knowledge.
   The evaluator's Level is predefined in this user's profile at the time of his registration into the community. The value of this "level" is a coefficient that depends on the significance of the evaluator's level with respect to all the existing levels in the community.

$$Coefficient(Level) = Score/N \quad (1)$$

   where: N is the number of levels

Table 1. Computation of the level of the evaluator

Level	Coefficient
Professor	5/5=1
PhD	4/5=0.8
Master's	3/5=0.6
Bachelor's	2/5=0.4
Basic	1/5=0.2

2. ***Score*:** For each evaluation, a score will be given by the evaluating member. This score represents the member's opinion of the solution being evaluated: the evaluator may assign a high score if the solution is good or very good, and an average score if the solution is not quite correct, or even a low score if the solution is judged incorrect.

3. ***Confidence:*** The score given by an evaluating member reflects his opinion. A percentage of confidence is assigned by the evaluator himself to each score he gives: if he is sure of his score, he will give a high degree of confidence; otherwise, the degree of confidence will be lower.

4. ***Evaluation context:*** In evaluating a solution, the member does his assessment of the solution based on a given source. This source could be a book, a document, an article, etc. The importance of the sources is different based on each one's credibility. We thus assign a weight to each source to indicate its reliability. The evaluation context can take several formats: tested results, research outcome, a similar problem, an approved opinion, or a new problem.

5. ***Skill level:*** A profile is associated with each member of the community. This profile is mainly used to retrieve information about the expertise of a member according to what he/she described as areas of expertise with respect to all the areas identified in the community by the input parameter "degree of expertise". As such, for a given problem in a specific domain, a member will have a certain level of expertise that we call "skill level". More precisely, this level represents the quality of a member in relation to his expertise in a specific area. This will allow us to get an idea about the credibility of his evaluation.

   Three cases can be distinguished:

   - Either the domain belongs to the member's skills set (high rate)
   - Or the domain belongs to the member's centre of interest (average rate)
   - Or the member has no knowledge at all about the domain (low rate)

   In order to calculate the skills level of a member with respect to the domain of the solution under evaluation, we need to calculate the similarity between all of the member's areas of expertise and the domain of this solution. To this end, we have defined a taxonomy that encompasses the existing domains in a given community and have added rules that cover all the possible cases for the relative positions of two domains in the taxonomy. We summarize these rules in the following table:

Table 2. Similarity rules

	Description	**Similarity**
Rule 1	D is the same as D'	Sim = 1
Rule 2	D (direct or indirect) parent of D'	Sim = 1
Rule 3	D (direct/indirect) son of D'	$\text{Sim} = \frac{\lvert \text{weight}(D) - \text{weight}(D') \rvert}{\Delta \text{Lev}}$
Rule 4	D and D' are independent	$\text{Sim} = \begin{cases} \text{weight}(D_c) \text{ if D and D' at the same level} \\ \frac{\text{weight}(D_c)}{\Delta \text{Lev}} \text{ otherwise} \end{cases}$

where:

D is the problem domain;
D' is one of the member's domains of expertise
Dc is the closest (parent) domain common to D and D'
ΔLev is the difference between the levels of D and D'

The skill level of a member is calculated in relation to all his areas of expertise using an ontology that uses the mentioned degrees of expertise (see Fig.2). These degrees reflect the coverage of subdomains (son nodes) from the related domain (parent node) in terms of knowledge. For example, a member who has knowledge in the field IGraphic, *a-fortiori* covers the subdomains Swing and AWT.

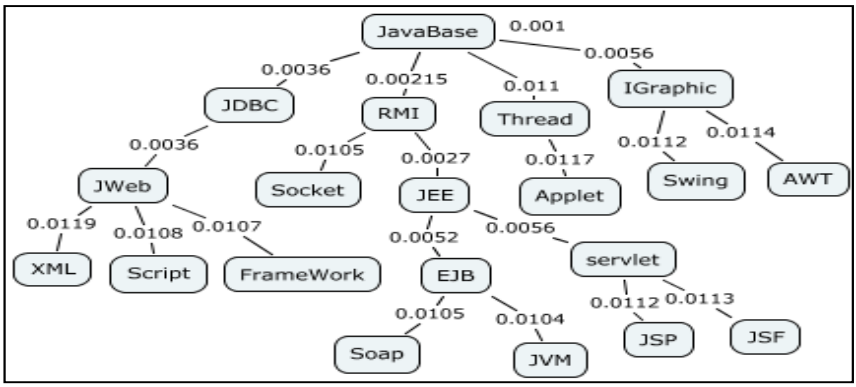


Fig. 2. Example of a taxonomy for the Java domain

This skill level is calculated according to the following steps:
Calculate the similarity between each domain of expertise D_i and the desired/searched-for domain D:

$$Similarity(D, Di) \tag{2}$$

Associate the similarity of each domain with the member's degree of expertise on D_i

$$Similarity(D, Di) * Expertise(M, Di) \tag{3}$$

Find the domain that corresponds to the value that maximizes the obtained values using the expression:

$$k = argmax_i(Similarity(D, Di) * Expertise(M, Di)) \tag{4}$$

Calculate the *quantum* which represents the amount of acquired skills in other domains to be added to the global competence of the member. The aim is not to neglect this additional skill.

$$Quantum= \{\{i = 1 - N\} \text{ and } i \neq k, (Similarity(D, Di) * Expertise(M, Di) / 10*(N - 1)\} \tag{5}$$

We point out that the number 10 is a factor that we can manipulate to increase or decrease the quantum with which we will adjust the additional competency provided by other domains rather than that giving the maximum of similarity.
Calculate the final competency of M with respect to D given that this value must be at least equal to a maximum value and does not exceed the value 1.

$$Competency\ (M, D) = min[1, ((S_k * P_k) * (1 + quantum))] \tag{6}$$

where:

S_k is the similarity between the domain D_k which maximizes the competence of M and the domain of the solution
P_K is the member's expertise
D is the domain of the solution

6. The degree of success: An active member may be considered as a trusted source because of his correct assessments in two cases: (1) he generally evaluates positively solutions that at the end of the validation process obtain a high degree of validity; and (2) he generally evaluates negatively solutions that at the end of the validation process are assigned a low degree of validity.
Accordingly, we assign to him a degree of success which represents the distance of one of his evaluations to the final validity of the solution according to the score he gave it. The following is the formula we propose to calculate the degree of success:

$$Success(Si) = 1 - |Score - validity(Si)| \quad (7)$$

A member's success score is calculated with respect to the relative success of all the solutions he has evaluated:

$$Success = \frac{1}{N}\sum_{i=1}^{N}(Success(Si)) \quad (8)$$

3.2 Use of Machine Learning Techniques

Machine Learning (ML) techniques are powerful in terms of their flexibility and ease of extraction of hidden relations that exist within data of the various applications they could be used for. We have decided to use ML in our problem of social validation of solutions; the intuition is to have automated learning from past experience of users' evaluations and the experts' assessments of the quality of these evaluations. We use the representation of an evaluation as presented above and apply different ML techniques on data given in this representation.

3.2.1 Modelling of Machine Learning Methods

We are mainly interested in three methods: Genetic Algorithms (GA) (See (Goldberg, 1989), (Holland, 1992) and (Mitchell, 1996)); Neural Networks (NN) (See (Muller and Reinhardt, 1994) and (Fausett, 1994)); and the Naïve Bayes Approach (Mitchell, 1997), as presented in the following sub-sections:

3.*2.1.1 Genetic Algorithms*: Genetic algorithms (GA) are often used for optimization. In our case, we have used a similar approach to Data Mining guided by the GA to highlight useful information for solving our problem. The approach proceeds as follows in the learning phase:

- Consider an initial population as a set of evaluations, carried out on different solutions, and represented using the six aforementioned parameters.

- Conduct a series of crossings and mutations by randomly changing the parameter values.
- Keep the final population which corresponds to the final validity predicted by a supervisor. The fitness function used is the following:

$$Fitness = \frac{1}{N+1}\sum_{i=1}^{N}(c_i * x_i) \quad (9)$$

where x_i is the value of the attribute and c_i its coefficient.

- Encode the population obtained after the previous step using a binary encoding.
- Apply an algorithm for mining association rules such as the Close algorithm (Pasquier et al., 1999; Pei et al, 2000).

Gradually enrich the rule base with new rules. These new rules will be added to cater for newly encountered cases that are not covered by the already generated rule base.

In our context, an Association Rule (Han et al., 2006; Sarawagi et al, 2000) is an implication of the form X → Y where X is a conjunction of Attribute-Value pairs of the form "$Attribute_i = Value_j$" and Y is a pair Attribute-Value of the same form which represents the degree of validity of an evaluation. Example:

$$If\ \big((Score = 0{,}8)\ and\ (Confidence = 0{,}6) and\ (Success = 0{,}9)\big)\ then\ Validity = 0{,}8 \quad (10)$$

The CLOSE algorithm is used for the extraction of informative association rules based on the informative content of the database (Pasquier et al., 1999; Pei et al, 2000).

3.2.1.2 *Neural Networks*: Artificial Neural Networks (ANNs) have been designed to mimic information flow in the human brain. Neural networks are efficient for complex problem solving, especially for pattern matching, classification, and optimization problems. In our case, we have used this method to predict a degree of validity of a given evaluation.

After designing several models of ANNs, we have tried various learning and activation functions, varying the number of neurons in each case. We have selected the NN architecture as follows for as good a learning phase as possible:

- Design a multilayer perceptron containing: six inputs, twenty neurons on the hidden layer and one neuron on the output layer.
- Feed in input into the network each 6-value input describing an evaluation of a solution.
- Give as output the validity score given by the supervisor for the given input.
- Train the network until the best learning is obtained (trying various architectures).
- Once a good learning has been achieved, simulate the network.

3.2.1.3 Naive Bayes Approach: The Naive Bayes Approach is a probabilistic approach based on conditional probability calculations. It is called Naïve due to the assumption it makes of independence of the various events (attributes) it considers. In spite of this, this assumption has not prevented them from providing an efficient and often good approach. In our case, we have considered a set of evaluations on various solutions, represented using the six parameters. We present below the steps followed for the calculation of the probabilities:

- Calculate the probability of the validity value Vi :

$$P(V_i) = \frac{Frequency(value_i)}{N} \tag{11}$$

where N is the number of evaluations considered for the learning.

- Calculate the conditional probability that an attribute takes a value $value_i$, given that V_i is the value of the validity:

$$P(value_i/V_i) = \frac{Frequency(value_i)}{P(V_i)} \tag{12}$$

- Calculate the conditional probabilities that the validity takes different possible values V_i bearing in mind that the attributes have some given values.

$$P(Vi/attribute_i = value_1, attribute_j = value_2, \ldots, attribute_n = value_{\mathrm{m}}) \tag{13}$$

- Consider the validity V_i corresponding to the maximal probability of the set obtained as the validity of the evaluation.

3.2.2 Application of the Machine Learning Techniques

In the case of a new solution to be validated, the steps to be followed are as follows:

- Represent all the evaluations of the solution using the six attributes.
- Predict a degree of validity for each one of the evaluations for the three methods.
- In the case of the GA: associate with each evaluation the most suitable rule, and then generate rules based on the validity rates for each evaluation.
- In the case of the ANN approach: simulate each evaluation by the ANN assigning a validity rate for each evaluation.
- In the case of the Naïve Bayes approach: calculate the probabilities of the evaluations and associate each evaluation with a validity rate.
- Remove the incorrect values from the set of validity rates.
- Apply a credibility formula to calculate the final validity:

$$FinalValidity = \sum_{i=1}^{N} (Credibility_i * Validity_i) \tag{14}$$

4 Implementation and Tests

4.1 Description of the Testing Phase

In order to test our approach, we have developed an online community platform for Java learning, including the different functionalities related to our approach. Furthermore, we have developed a prototype to automatically generate a large number of solutions and their evaluations. This has allowed us to create the database and hence to carry out our learning process. In addition, we have used the community platform, where members can add new problems, propose new solutions or evaluate some existing ones. Then we have represented all the obtained evaluations according to the six parameters, as proposed in section III. Finally, we have applied the selected ML techniques.

4.2 Discussion of the Findings

We have implemented the three ML techniques to predict the degree of validity of a given solution. After some tests and experimentations, we obtained the results shown in the following figures:

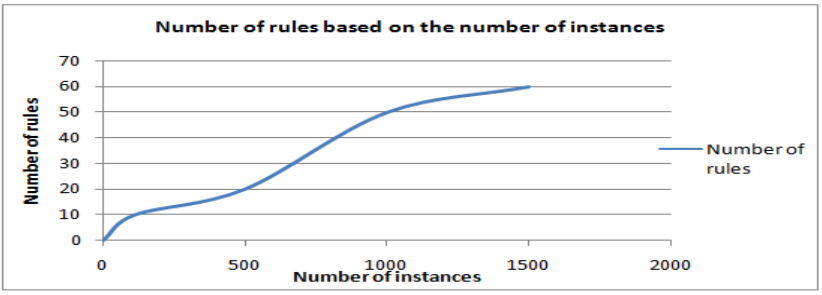


Fig. 3. Increasing number of the association rules (GA approach)

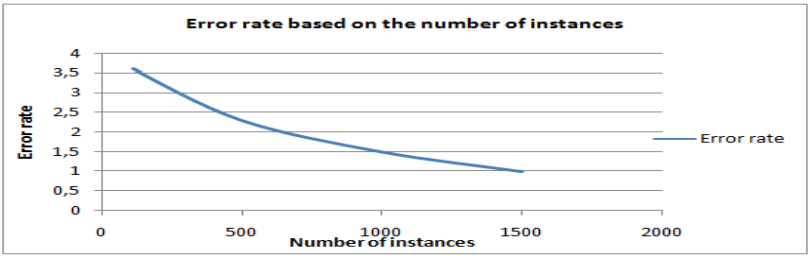


Fig. 4. Error rate of the ANN method

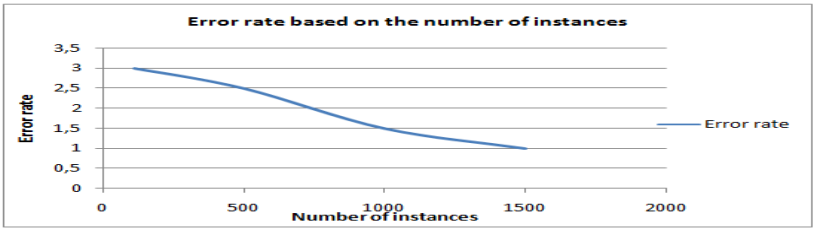


Fig. 5. Error rate of the Naïve Bayes approach

- We can deduce from the application of the GA approach that more and more new solutions are validated, more rules can be generated and the accuracy of the rules application increases improving the degree of validity (see Fig. 3).
- The application of ANNs allows us to conclude that the error rate decreases with the increasing number of examples that are used as input during the ANN learning phase (see Fig. 4).
- Finally, the application of the Naïve Bayes approach allows us to deduce that with the increasing number of examples, the error rate decreases, which implies that the number of correct predictions increases, and hence the results become more accurate (see Fig. 5).

On the other hand, an analysis of the results we have obtained has allowed us to compare the three ML methods on the basis of four criteria (see Table 3).

- *Criterion 1 – Data redundancy:* if data appears frequently, then the learning outcomes are improved.
- *Criterion 2 – Number of instances:* when the number of data instances used in the learning increases, the learning is better guided and gives more accurate results.
- *Criterion 3 – Appropriateness of results:* the learning is considered good if it frequently gives accurate results with low error rates.
- *Criterion 4 –New cases: l*earning is considered good if it can handle well and robustly a situation where a new case arises (a case which was not seen during the learning phase).

Table 3. Comparison of the three ML methods

	GA	ANN	NBA
Criterion 1	Yes	No	Yes
Criterion 2	Yes	Yes	Yes
Criterion 3	Somewhat	Strongly	Always
Criterion 4	Frequently	Somewhat	Never

According to these results and analysis, we conclude that the neural networks have given the best performance compared to the two other approaches. In order to improve the obtained results, it would be very interesting to combine these approaches in different ways and to compare the performance of the different algorithms.

5 Conclusion and Perspectives

We are interested in this work in the problem of social validation of solutions proposed in the context of a learning community. We have considered the evaluations carried out on already proposed solutions and modeled the problem according to several criteria. An automatic validation process was proposed using three machine learning techniques: genetic algorithms, neural networks and the Naive Bayes Approach. An experimental study of the developed prototype has been conducted. The results show that neural networks have given the best performance.

As future work we envisage to make further tests on a real community of learners and collect as much data as possible to enrich the learning. We envisage also to check the possibility of combining some of the different learning techniques and to generalize the process of social validation of more than one proposed solutions to the same problem.

References

1. Berkani, L., Driff, L.N., Guessoum, A.: Social Validation of Learning Objects in Online Communities of Practice Using Semantic and Machine Learning Techniques. In: Amine, A., Mohamed, O.A., Bellatreche, L. (eds.) Modeling Approaches and Algorithms. SCI, vol. 488, pp. 237–247. Springer, Heidelberg (2013)
2. Cabanac, G., Chevalier, M., Chrisment, C., Julien, C.: Social validation of collective annotations: Definition and experiment. Journal of the American Society for Information Science and Technology 61(2), 271–287 (2010)
3. Fausett, L.: Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Prentice Hall (1994) ISBN: 0133341860
4. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Professional, Reading (1989), ISBN 978-0201157673
5. Herr, K., Anderson, L.G.: The Action Research Dissertation: A Guide for Students and Faculty. Thousand Oaks (2008) ISBN 0-7619-2991-6
6. Holland, J.: Adaptation in Natural and Artificial Systems. MIT Press, MA (1992)
7. Laister, J., Kober, S.: Technikum Joanneum Social Aspects of Collaborative Learning in Virtual Learning Environments, http://comma.doc.ic.ac.uk/inverse/papers/patras/
8. Le Boulch, D., Bouyssou, D., Grundstein, M.: Towards a redefinition of the relationships between information systems development and individual cognition. In: Information Technologies in Environmental Engineering, Springer, Heidelberg (2009)
9. Mitchell, T.: Machine Learnin. McGraw Hill (1997) ISBN 007042807
10. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, MA (1996)
11. Muller, B., Reinhardt, J.: Neural Networks. Springer (1991)
12. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: 7th International Conference on Database Theory (January 1999)
13. Pei, J., Han, J., Mao, R.: Closet: An effcient algorithm for mining frequent closed itemsets. In: SIGMOD Int. Workshop on Data Mining and Knowledge Discovery (2000)
14. Swan, K., Shea, P.: The development of virtual learning communities. In: Hiltz, S.R., Goldman, R. (eds.) Asynchronous Learning Networks: The Research Frontier, pp. 239–260. Hampton Press, New York (2005)
15. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers (March 2006) ISBN 1-55860-901-6
16. Sarawagi, S., Thomas, S., Agrawal, R.: Integrating Association Rule Mining with Databases: Alternatives and Implications. Data Mining and Knowledge Discovery Journal 4(2/3) (2000)

Remotely Sensed Data Clustering Using K-Harmonic Means Algorithm and Cluster Validity Index

Habib Mahi[1(✉)], Nezha Farhi[1], and Kaouter Labed[2]

[1] Earth Observation Division, Centre of Space Techniques, Arzew, Algeria
[2] Kaouter LABED, Faculty of Mathematics and Computer Science Mohamed Boudiaf, University – USTOMB, Oran, Algeria
{hmahi,nfarhi}@cts.asal.dz,
kaouther.labed@univ-usto.dz

Abstract. In this paper, we propose a new clustering method based on the combination of K-harmonic means (KHM) clustering algorithm and cluster validity index for remotely sensed data clustering. The KHM is essentially insensitive to the initialization of the centers. In addition, cluster validity index is introduced to determine the optimal number of clusters in the data studied. Four cluster validity indices were compared in this work namely, DB index, XB index, PBMF index, WB-index and a new index has been deduced namely, WXI. The Experimental results and comparison with both K-means (KM) and fuzzy C-means (FCM) algorithms confirm the effectiveness of the proposed methodology.

Keywords: Clustering · KHM · Cluster validity indices · Remotely sensed data · K-means · FCM

1 Introduction

Clustering is an exploratory data analysis tool that reveals associations, patterns, relationships, and structures in masses of data [1] [2]. Two approaches of clustering algorithms exist in the literature: fuzzy (or soft) and crisp (or hard) clustering. In the first approach, clusters are overlapping and each object belongs to each cluster to a certain degree (or with a certain fuzzy membership level) [3]. The fuzzy c-means (FCM) [4] seems to be the most widely used algorithm in the field of fuzzy clustering. It appears to be an appropriate choice in multiple domains as remote sensing satellite images and pattern recognition [5] [6]. In crisp clustering, clusters are disjoint: each objet belongs to exactly one cluster as example we cite the K-Means (KM) [7] and ISODATA (Iterative Self-Organizing Data Analysis Technique) algorithms [8]. These latter are widely used clustering methods for multispectral image analysis [9]. Also, these algorithms have been successfully used in various topics, including computer vision and astronomy. Their popularity is mainly due to their scalability and simplicity. However, they suffer from a number of limitations. Firstly, the requirement to define a priori the number of K clusters is considered as a handicap and consequently an inappropriate choice of initial clusters may generate poor clustering results [10]. Secondly,

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 105–116, 2015.
DOI: 10.1007/978-3-319-19578-0_9

the KM algorithm and similarly the ISODATA algorithm work best for images with clusters which are spherical and that have the same variance. This is often not true for remotely sensed data with clusters which are more or less elongated with a much larger variability, such as forest for example [11]. Also, convergence to local optimum is always observed in this kind of algorithms [1].

To deal with these drawbacks, considerable efforts have been made to mainly create variants from the original methods. As examples we cite KM and its alternatives K-Harmonic Means, Trimmed k-means and k-modes algorithm [1]. At the same time some works have focused on the developing of measures to find the optimal number of clusters using cluster validity indices [3]. We distinguish fuzzy indices (used with fuzzy clustering) and crisp indices (used with hard clustering). As examples of fuzzy indices we can mention XB index [12] as well as Bezdek's PE and PC indices [13] [14]. DB-index [15], Dunn's index [16] and Calinski-Harabasz index [17] are some of the popular indices used in crisp clustering. [3] [18] [19] give a very important review of different CVIs present in the literature.

In this study we investigate the ability of the K-Harmonic Means clustering algorithm combined with validity indices, especially in unsupervised classification of remote sensing data. The rest of paper is organized as follows. Methodology will be firstly presented in Section 2; the experimentation and the results obtained will be tackled in Section 3. Section 4 concludes the paper.

2 Methodology

In this section we give a brief description of the K-Harmonic Means and four clustering validity indices. Then we present the proposed method in details. In the next sections, the following notation will be adopted:

N: The number of objects in the data set.
x_i: The i^{th} object in the data set.
K: The number of clusters.
c_j: The center of cluster j.
d: The number of dataset dimensions.

2.1 K-Harmonic Means Algorithm

The initialization of centers influence on the K-Means (KM) performance and it is considered as the main drawback of this algorithm. To improve KM, Zhang [20] proposes to use the harmonic mean instead of standard mean in the objective function and has named the new algorithm K-Harmonic Means (KHM).

$$KHM = \sum_{i=1}^{N} \frac{K}{\sum_{j=1}^{K} \frac{1}{\left\| x_i - c_j \right\|^q}} \tag{1}$$

New centers clusters are calculated as following [21][22]:

$$c_k = \frac{\sum_{i=1}^{N} \frac{1}{\left[\sum_{l=1}^{K} \frac{\|x_i - c_j\|^q}{\|x_i - c_l\|^q}\right]^2} x_i}{\sum_{i=1}^{N} \frac{1}{\left[\sum_{l=1}^{K} \frac{\|x_i - c_j\|^q}{\|x_i - c_l\|^q}\right]^2}} \quad (2)$$

2.2 Clustering Validity Indices

In this sub-section, we introduce the clustering validity indices used in this work, namely Davies-Bouldin (DB), Xie-Benie (XB), Pakhira-Bandyopadhyay-Maulik Fuzzy (PBMF), WB index (WB) and WB-XB index (WXI).

- Davies-Bouldin index (DB ↓) [15]: It is a very popular and used crisp index in clustering algorithms. It requires only two parameters to be defined by the user, the distance measure noted p and the dispersion measure noted q. The DB is defined as follows:

$$DB = \frac{1}{K} \sum_{i=1}^{K} R_i \quad (3)$$

With

$$R_i = max_{i,i \neq j} \left\{ \frac{S_i + S_j}{M_{ij}} \right\} \quad (4)$$

Where

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} \left\| x_j - c_i \right\|^q \right\}^{\frac{1}{q}} \quad (5)$$

And

$$M_{ij} = \left\{ \sum_{k=1}^{K} \left\| c_{ki} - c_{kj} \right\|^p \right\}^{\frac{1}{p}} \quad (6)$$

With
c_{ki}: k^{th} Component of the n-dimensional vector c_i.
c_i: The center of cluster i.
M_{ij}: The Minkowski metric.
T_i: The number of vectors (pixels) in cluster i.

- Xie-Benie index (XB ↓) [12]: Also called function S, is defined as a ratio of the total variation to the minimum separation of clusters. Its definition is:

$$XB = \frac{1}{N} \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} (\mu_{ij})^m \|x_j - c_i\|^2}{min_{l \neq i} \|c_l - c_i\|^2} \quad (7)$$

- Pakhira-Bandyopadhyay-Maulik Fuzzy index (PBMF ↑) [3]: It is considered as validity index measure for fuzzy clusters. It is formulated as follows:

$$PBMF = \frac{1}{K} \times \frac{E_1}{\sum_{i=1}^{N} \sum_{j=1}^{K} (\mu_{ij})^m \|x_i - c_j\|^2} \times max_{l \neq i} \|c_l - c_i\|^2 \quad (8)$$

With E_1 is constant for a given dataset.

- WB index (WB ↓) [23]: It is defined as a ration of the measure of cluster compactness to its measure of separation. It is given by:

$$WB = K \frac{\sum_{i=1}^{N} \|x_i - c_{pi}\|^2}{\sum_{i=1}^{K} n_i \|c_i - \bar{X}\|^2} \quad (9)$$

- WB-XB index (WXI ↓): It is defined as the average between WB and XB indices and is formulated as follows:

$$\text{WXI} = (\text{WB\_index} + \text{XB\_index}) / 2 \quad (10)$$

2.3 Mean Square Error (MSE)

It is a measure of error which is often used in clustering problems. It represents the mean distance of objects in the dataset from the nearest centers [24]. It is formulated as follows:

$$MSE = \frac{\sum_{j=1}^{K} \sum_{X_i \in C_j} \left\| x_i - c_j \right\|}{N * d} \quad (11)$$

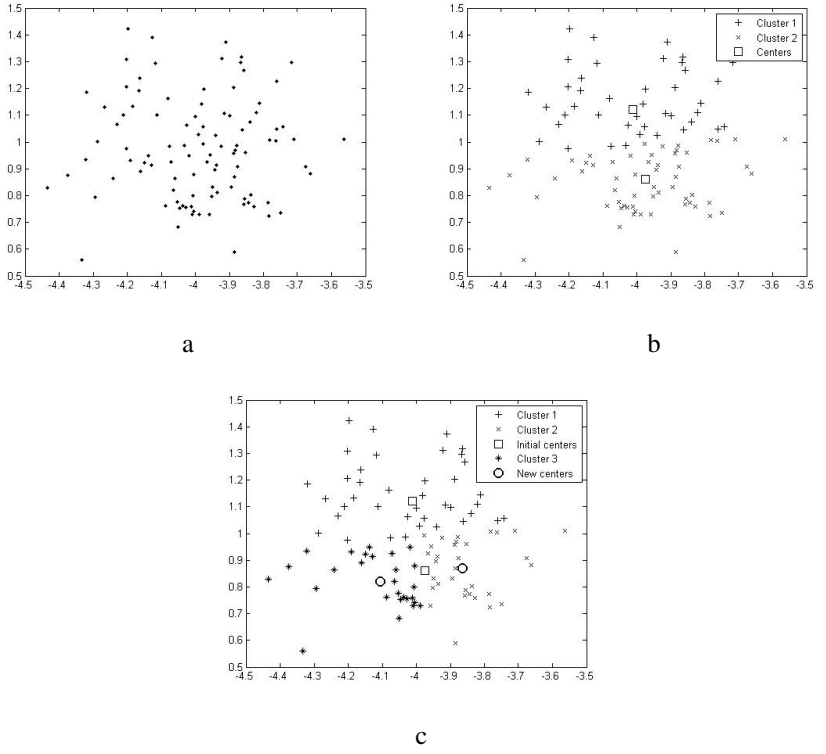
2.4 Proposed Method

In this subsection, we present the proposed method which combines the KHM algorithm, the mean square error (MSE) and WXI cluster validity index. This new method is called Growing KHM (GKHM).

For a given data distribution two centers are chosen randomly (Fig. 1. a), the KHM clustering algorithm is then applied to obtain the two initial clusters (Fig. 1. b). Also, the mean square error (MSE) is computed in this stage for each cluster to select the heterogeneous one (MSE is maximal) to be divided. Therefore, two new centers are computed (Fig. 1. c) and the old one is removed. The process is repeated until a number of epochs are satisfied. The complete algorithm for the proposed method is given by the following:

1. Choose two centers randomly from the dataset.
2. Run the KHM algorithm with these two centers
3. **Repeat**

4. epoch =1
5. Compute MSE for each cluster
6. Select the cluster with the maximum MSE value
    - Insert two new centers halfway between the old center and the two extremes of the cluster in order two have two new clusters.
    - Remove the old center
    - Run the KHM algorithm with the new centers (K = 2).
7. Compute the WXI^{epoch} of all the clusters and save it in the vector V with the related centers
8. **Until** (epoch number's reached)
9. Select the minimum value of WXI in V i.e. The final number of clusters
10. Clustering dataset with the appropriate centers.



a b c

Fig. 1. Process of new centers: a) Data Initialization, b) Data Sampling, c) New Centers Generation

3 Experimental Results

This section is devoted to experiments that ensure the validity and effectiveness of the proposed method. It is divided into three subsections. In the first subsection, an experiment is conducted by using synthetic datasets to select the most suitable cluster validity

index for our work. In the second subsection, a comparison of our approach with both KM and FCM algorithms is drawn. The last subsection concerns the clustering of real satellite images using the proposed method and its results. All the experiments results have been obtained using the MATLAB software package.

3.1 Comparison between the Four Cluster Validity Indices

In order to select the best clustering validity index, we experimentally evaluated their performance on four different synthetic datasets using the basic KHM. Some of the datasets namely S1 and S4 are plotted in Fig. 2 respectively. Each dataset consists of 5000 points representing 15 clusters. All the datasets can be found in the SIPU web page `http://cs.uef.fi/sipu/datasets`.

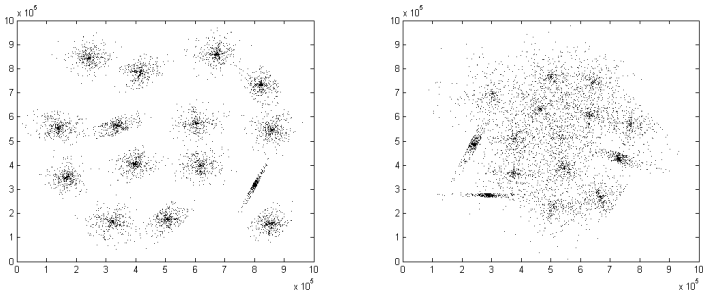


Fig. 2. Synthetic data S1 and S4

In this paper, we have used four synthetic datasets S1-S4 which have the same number of clusters (K=15) and the same Gaussian distribution with increasing overlap between the clusters. The overlapping is an additional criterion which allows us to select the optimal cluster validity index between indices used in this work. For this end, we have applied the KHM algorithm as mentioned before for each dataset by varying the number of clusters from 2 to 20; and, the values of the four CVI's are computed for different K. The results reported in Tables 1 and 2 show only the best values obtained by the four CVI's and their corresponding number of clusters K.

From Table 1, we can see that WB and XB cluster validity indices give the best values for the KHM algorithm and reach their minimum respectively at the optimal

Table 1. Comparison between DB, WB, PBMF and XB indices for S1 dataset

	Cluster Validity Indices			
K	DB	WB	PBMF	XB
13	**0.40**	0,49	2,99 x 10^{10}	0,08
14	0.42	0,36	**9,69 x 10^{10}**	0,06
15	0.44	**0,24**	4,26 x 10^{10}	**0,04**

Table 2. Comparison between DB, WB, PBMF and XB indices for S4 dataset

	Cluster Validity Indices			
K	**DB**	**WB**	**PBMF**	**XB**
4	0.84	1.89	**2.32 x 10^{10}**	0.16
11	**0.64**	1.17	1.18 x 10^{10}	0.11
14	0.65	0.96	1.25 x 10^{10}	**0.09**
15	0.72	**0.90**	0.77 x 10^{10}	0.14

number of clusters (K=15). On the other hand, the DB and PBMF cluster validity indices approximate the number of clusters (K=13 and K=14).

From Table 2, we notice that WB index still offers the best values for the KHM algorithm and reaches its minimum for the optimal clusters number (K=15). The XB index approximate the solution and has its optimum value nearly to the solution (K=14). However, DB and PBMF fail to find a near best solution by returning a completely wrong number of clusters (K=11 and K=4) and having an unstable minimum.

In Summary, the results show that all the cluster validity indices provide an accurate estimation of the clusters number when the clusters in dataset present a small distortion. However, several knee points are detected with exception of the WB index. For clusters with a largest distortion, case of the S3 and S4 datasets, the DB and PBMF indices fail to find the optimal number of clusters. These conclusions lead us to say that only the WB and the XB indices can be used for this kind of datasets. According to the obtained results; the combination of WB and XB indices seems interesting and the new index called WXI (Equation 9) was deduced and tested.

Table 3. Comparison of the minimal values of the WXI for S1,S2,S3 and S4

	S1	**S2**	**S3**	**S4**
OVI	0.13	0.23	0.45	0.49
K	15	15	14	15

The results of WXI are very promising since the error margin reported in Table 3 is acceptable. Indeed, the combination of WB and XB indices has maximized the performances of both of them and erased the deficiency of each one.

3.2 Comparison with KM and FCM Algorithm

In this section, different tests have been performed using GKHM, KM and FCM over 50 iterations. The WXI has been computed in each test and used to compare between their results.

In order to compare between the GKHM and the two other algorithms using the WXI, we have computed up each of them to 50 iterations with a static number of clusters (K=15) for the KM and the FCM. The results appear in Figures 3 and 4.

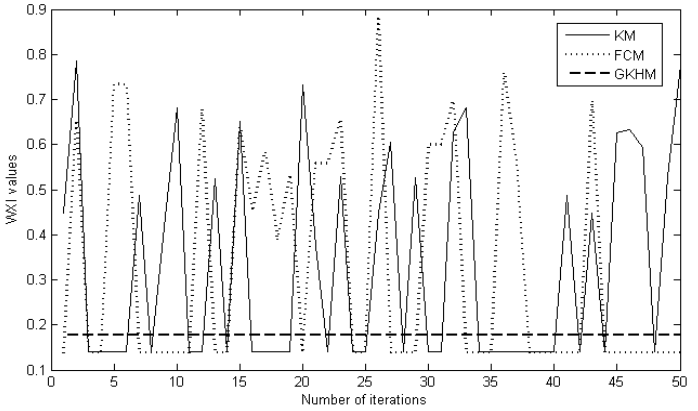


Fig. 3. Comparison between the GKHM, the KM and the FCM for S1 using the WXI

Form Fig. 3, we notice that the KM and the FCM reach inferior minimums than the GKHM but the results are very fluctuant and change constantly; it brutally increases after reaching the minimum which indicates unstable algorithms unlike the GKHM which is totally stable and remains on its minimum value.

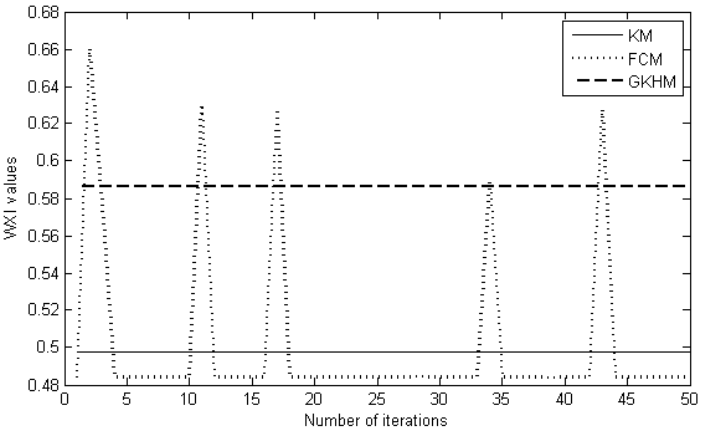


Fig. 4. Comparison between the GKHM, the KM and the FCM for S4 using the WXI

The results in Fig. 5 represent the WXI values for the high overlap data synthetic S4. The curves shape shows that the KM and GKHM are stable; however, only the first algorithm gives the best results. In contrast, the shape of the FCM curve stays very fluctuant and unstable.

Table 4. The average results of the WXI at 50 iterations for all synthetic datasets

	Aver-age_WXI_KM	Average_ WXI _FCM	Average_ WXI _GKHM
S1	0.45	0.43	**0.17**
S2	0.32	0.30	**0.29**
S3	**0.42**	0.45	0.53
S4	**0.49**	0.50	0.64
Average	0.42	0.42	**0.40**

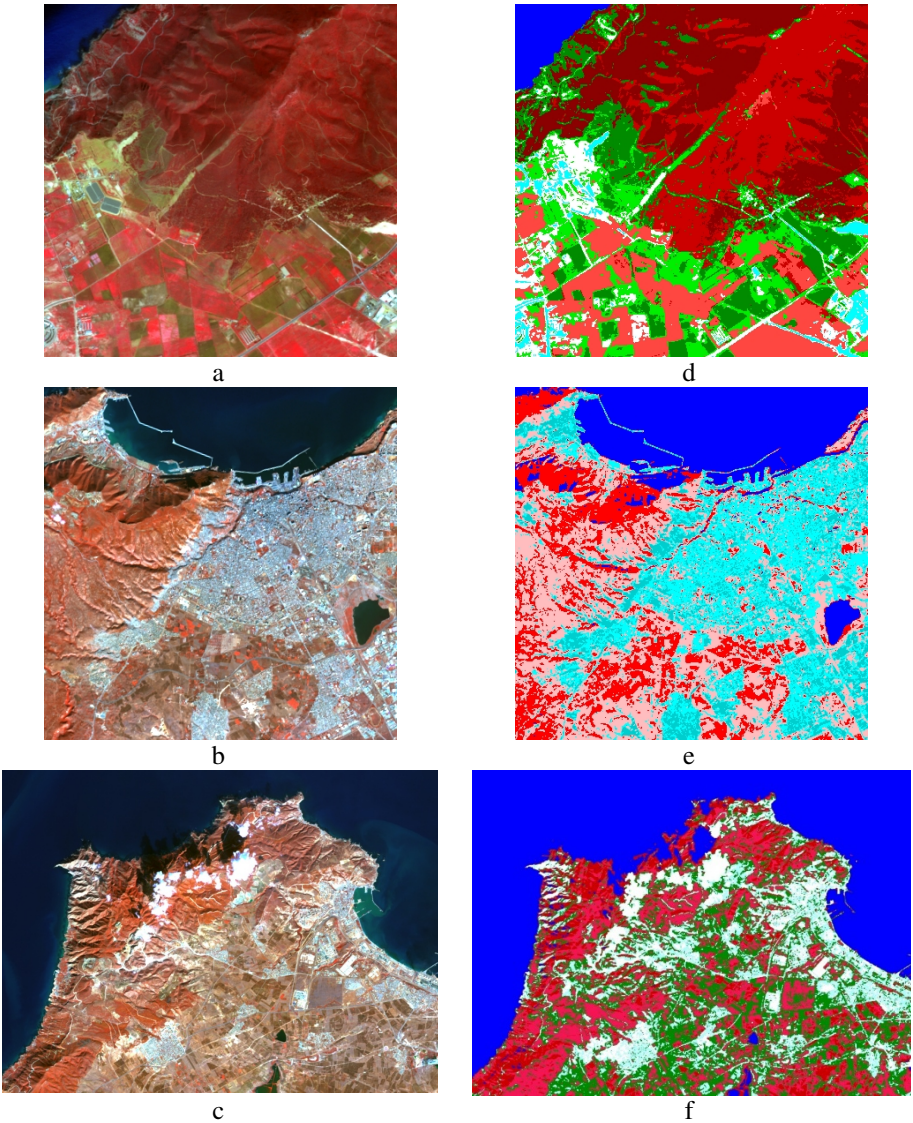


Fig. 5. Clustering using the GKHM on remote sensed data sets

From Table 4, we notice that the GKHM is a totally stable algorithm and tends to minimize the WXI values more than the KM and FCM, especially when data are well-separated. However, the GKHM responds less well when dealing with high overlapped datasets. In the case of FCM and KM, the results are unstable due to their high dependency on their centers number initialization. The three algorithms have approximately the same results with better global issues for GKHM concerning datasets tested in this paper.

3.3 Experiment on Remotely Sensed Data

In the last experiment, the clustering has been performed on three multispectral remotely sensed data; the details of the image sets are as follows:

- A Landsat 8 sub-scene of Oran the image has three spectral data channels and size of 400 x 400. The spatial resolution is 30 meters (Fig. 6.a).
- A Spot 5 sub-scene of Oran the image has three spectral data channels and size of 400 x 400. The spatial resolution is 20 meters (Fig. 6.b).
- A Landsat 8 sub-scene of Arzew the image has three spectral data channels and size of 600 x 800. The spatial resolution is 30 meters (Fig. 6.c).

The clustering results of the three remotely sensed data by the proposed method are shown in Fig. 6.d with eight clusters, Fig. 6.e with five clusters and Fig. 6.f with six clusters, respectively. The visual comparison with the corresponding original images shows that the obtained results appear generally satisfying even if we notice some confusion between water pixels and shadow ones, case of the second image.

4 Conclusion

A new clustering method for multispectral remotely sensed data has been proposed in this paper. The method combines both the K-Harmonic means algorithm and the clustering validity index in order to find the optimal number of clusters and perform the classification task. Note that the K-harmonic means has been used with only two clusters and the increasing of the centers number has been provided by an automatic insertion of the new clusters. However, some improvements can be made, especially by reducing the time processing cycle. Also, the developed algorithm uses internally a combination of validity indices in order to return an optimal number of clusters.

Other improvements could be done by testing the GKHM on large datasets including high-dimensional datasets and shape sets.

A further research will involve the application of new validity indices such as DB* index [25], the comparison with both the enhanced differential evolution KHM [26] and the modified version of k-means algorithm proposed by Celebi and al. [27] and finally the use of the ensemble clustering technique.

References

1. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Serieson Statistics and Applied Probability. SIAM, Philadelphia (2007)
2. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood (1988)
3. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification. Fuzzy Sets and Systems 155, 191–214 (2005)
4. Bezdeck, J.C.: FCM: Fuzzy C-Means algorithm. Computers and Geoscience 10, 191–203 (1984)
5. Gong, X.-J., Ci, L.-L., Yao, K.-Z.: A FCM algorithm for remote-sensing image classification considering spatial relationship and its parallel implementation. In: International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR 2007, November 2-4, vol. 3, pp. 994–998 (2007)
6. Gao, Y., Wang, S., Liu, S.: Automatic Clustering Based on GA-FCM for Pattern Recognition. In: Second International Symposium on Computational Intelligence and Design, ISCID 2009, December 12-14, vol. 2, pp. 146–149 (2009)
7. McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symp. Mathematics, Statistics and Probability, pp. 281–296 (1967)
8. Ball, G., Hall, D.: ISODATA: A novel method of data analysis and pattern classification. In Technical report, Stanford Research Institute, Menlo Park, CA, USA (1965)
9. Huang, K.: A Synergistic Automatic Clustering Technique (Syneract) for Multispectral Image Analysis. Photogrammetric Engineering and Remote Sensing 1(1), 33–40 (2002)
10. Zhao, Q.: Cluster validity in clustering methods. Ph.D. dissertation. University of Eastern Finland (2012)
11. Korgaonkar, G.S., Sedamkar, R.R., KiranBhandari.: Hyperspectral Image Classification on Decision level fusion. In: IJCA Proceedings on International Conference and Workshop on Emerging Trends in Technology, vol. 7, pp. 1–9 (2012)
12. Xie, X.L., Beni, A.: Validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 3, 841–846 (1991)
13. Bezdek, J.C.: Cluster validity with fuzzy sets. J. Cybernet. 3, 58–73 (1974)
14. Bezdek, J.C.: Mathematical models for systematics and taxonomy. In: Eighth International Conference on Numerical Taxonomy, San Francisco, CA, pp. 143–165 (1975)
15. Davies, D., Bouldin, D.: A cluster separation measure. IEEE PAMI 1(2), 224–227 (1979)
16. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well separated clusters. J. Cybernet. 3, 32–57 (1973)
17. Calinski, R.B., Harabasz, J.: Adendrite method for cluster analysis. Commun. Statist. 1–27 (1974)
18. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Prez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition 46(1), 243–256 (2013)
19. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: Part II. SIGMOD Record 31(3), 19–27 (2002)
20. Zhang, B.: Generalized K-Harmonic Means Boosting in Unsupervised Learning. Technical Reports, Hewllet Laborotories, HPL-2000-137 (2000)
21. Zhang, L., Mao, L., Gong, H., Yang, H.: A K-harmonic Means Clustering Algorithm Based on Enhanced Differential Evolution. In: 2013 Fifth International Conference on Measuring Technology and Mechatronics Automation, 2014 Sixth International Conference on Measuring Technology and Mechatronics Automation, pp. 13–16 (2013)

22. Thangavel, K., Karthikeyani Visalakshi, K.: Ensemble based Distributed K- Harmonic Means Clustering. International Journal of Recent Trends in Engineering 2(1), 125–129 (2009)
23. Zhao, Q., Fränti, P.: WB-index: a sum-of-squares based index for cluster validity. Knowledge and Data Engineering 92, 77–89 (2014)
24. Malinen, M.I., Mariescu-Istodor, R., Fränti K-means*, P.: Clustering by gradual data transformation. Pattern Recognition 47(10), 3376–3386 (2014)
25. Thomas, J.C.R.: New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance. In: V Chilean Workshop on Pattern Recognition, November 11-15 (2013)
26. Zhang, L., Mao, L., Gong, H., Yang, H.: A K-harmonic Means Clustering Algorithm Based on Enhanced Differential Evolution. In: 2013 Fifth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), January 16-17, pp. 13–16 (2013), doi:10.1109/ICMTMA.2013.1
27. Emre, C.M., Kingravi, H.A., Vela, P.A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications (2013)

Computational Intelligence: BioInformatics

Comparison of Automatic Seed Generation Methods for Breast Tumor Detection Using Region Growing Technique

Ahlem Melouah(✉)

Department of Informatics, Labo LRI, Badji-Mokhtar Annaba University,
P.O.Box 12, 23000, Annaba, Algeria
ahlem.melouah@univ-annaba.dz

Abstract. Seeded Region Growing algorithm is observed to be successfully implemented as a segmentation technique of medical images. This algorithm starts by selecting a seed point and, growing seed area through the exploitation of the fact that pixels which are close to each other have similar features. To improve the accuracy and effectiveness of region growing segmentation, some works tend to automate seed selection step. In this paper, we present a comparative study of two automatic seed selection methods for breast tumor detection using seeded region growing segmentation. The first method is based on thresholding technique and the second method is based on features similarity. Each method is applied on two modalities of breast digital images. Our results show that seed selection method based on thresholding technique is better than seed selection method based on features similarity.

Keywords: Medical image segmentation · Medical informatics · Automatic seed selection · Region growing · Tumor detection

1 Introduction

The basic segmentation aim is to divide an image into different regions based on certain criteria. The regions with connected pixels of similar values can provide important cues for extracting semantic objects. Since, image segmentation is mainly used to locate an objects or an object boundary in an image thus it can be used in applications which involve a particular kind of object recognition such as breast tumor.

Though researchers introduced several images segmentation methods but, most of these methods are not suitable for medical images. Image segmentation using seeded region growing (SRG) technique has increasingly become a popular method because of its ability to involve a high-level knowledge of anatomical structures in seed selection process [Jianping et al. 2005]. In most of the region growing algorithms, all the neighbors need to be evaluated for the region to be grown. The region growing starts with a seed pixel and repeatedly adds new pixels as long as the segmentation criterion is satisfied [Deboeverie et al. 2013].

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 119–128, 2015.
DOI: 10.1007/978-3-319-19578-0_10

One of the most important factors in region growing process is seed pixel selection. Seed pixel is often chosen close to the center of the region of interest (ROI). For example, if we are to segment a tumor from the background, it is always advisable to select the seed point for the tumor in the middle of the tumor [Najarian and Splinter 2012]. If seeds are not properly selected, the final segmentation results would be definitely incorrect [Massich et al. 2011]. Despite the existence of many automatic seed selection methods, SRG algorithm still suffers from the problems of automatic seed generation [Mehnert and Jackway1997; Jianping et al. 2001].

In this paper, two automatic seed point selection methods are compared. The first method based on thresholding technique is proposed by Al-Faris et al. [Al-Faris et al. 2014]. The second method based on features similarity is proposed by Yuvaria and Ragupathy [Yuvarai and Ragupathy 2013]. The same data and the same criteria have been used in this comparison.

The rest of the paper is organized as follows: Section 2 describes experimental automatic seed selection methods. Section 3 gives a view on experimentation. Section 4 presents some results with discussion. Section 5 draws our conclusion.

2 Automatic Seed Selection Methods

For the region growing to be effectively achieved, the crucial part is the position of the seed pixel which must be selected from where the region growing may start [Mesanovic et al. 2013]. Up to now, some works use a semi-automatic region growing algorithm and still need user interaction for seed selection. Other works are fully automatic and the user has only a verification role. Among these later works those proposed by Al-Faris et al. [Al-Faris et al. 2014] and Yuvarai et al. [Yuvarai and Ragupathy 2013]. Al-Faris et al. exposed an automatic seed selection method based on the thresholding technique. Yuvaria et al. developed an automatic seed selection method based on features similarity. The description of these two methods is detailed in the following.

2.1 Seed Selection Method Based on Features Similarity (SSFS)

In order to detect a mass in a mammogram using SRG segmentation, Yuvarai and Ragupathy proposed a new seed point selection method based on features similarity. Statistical features like mean, dissimilarity, sum average, sum variance and auto correlation are considered as significant features able to identify a mass. These features are computed and fixed for masses which have been previously identified by an expert. Seed selection process starts by initializing a mask, and then calculates its features from the regions within the mask. If the mask features do not match with the mass predefined features, the mask is therefore shift. Otherwise, the initial pixel of the mask is taken as seed point.

2.2 Seed Selection Method Based on Thresholding Technique (SSTT)

Al-Faris et al. [Al-Faris et al. 2014] used SRG for breast MRI tumor segmentation with seed point selection based on the thresholding technique. A new algorithm is developed for automatic evaluation of the suitable threshold value. This algorithm searches for the maximum value in each row in the image and saves it temporarily. This process is repeated for all the rows until the last. Then, a summation of the temporarily stored values is calculated. The mean maximum raw is then calculated by dividing the summation value by the number of rows in the image. The resultant mean value will be considered as the threshold value for the binarization process. In order to remove the unwanted small white speckles in the image which do not belong to the ROI and enhance the boundary of the suspected regions, the morphological open operation (erosion followed by dilation operations) has been applied. To extract ROI, all the regions are ranked in an ascending order according to their density values. After, the highest region will be chosen as the main suspected region. The seed is the pixel of this main suspected region with maximum intensity value.

3 Experiments

3.1 Dataset

In this study, two databases with different modalities of breast digital images are considered:

1. RIDER breast MRI dataset downloaded from the National Biomedical Imaging Archive [10]. The dataset includes more than 1000 breast MRI images for five patients. All the images are axial 288 X 288 pixels. The dataset also includes Ground Truth (GT) segmentation which has been manually identified by a radiologist.
2. MiniMIAS database provided by the Mammographic Image Analysis Society (MIAS) [11]. MiniMIAS consists of a variety of normal mammograms as well as mammograms with different characteristics and several abnormalities. The mammograms are digitized at a resolution of 1024x1024 pixels and at 8-bit grey scale level. All the images include the locations of all the abnormalities that may be present.

3.2 Seed Point Selection Criterion

Region growing is one of the most popular techniques for medical images segmentation due to its simplicity and good performance [Saad et al. 2012]. But, this performance is deeply influenced by seed point position. Therefore, selecting a good set of initial seeds is very important. To determine the good seed position, Massich [Massich et al. 2011] tested 10 areas-of-interest selected at different distances and orientations from the lesion center. The 10 tested areas are: The area 1 is the zone located outside the lesion; the areas from 2 to 5 are the zones situated on the boundaries of the lesion; the areas from 6 to 9 are the zones placed near the lesion center and, the area 10 is the lesion center. The best segmentation results are obtained by using the seed

points located in area 10. The segmentation performance decrease when the seed position moves away from the lesion center. Consequently, a seed point can be placed in three different areas:

1. Inside the ROI; in this situation, segmentation result is more and more accurate if seed position approximates the ROI center.
2. On the border of the ROI; in this situation, there are two possibilities, either the segmentation fails or success.
3. Outside the ROI; in this situation, the segmentation fails.

Figure 1 gives an example of these three situations. If the seed is centered in the ROI (figure 1.a), therefore the SRG segmentation well extracts the lesion (figure 1.e). If the seed is placed on the border of the ROI (Figure 1.b and Figure 1.c), therefore the SRG segmentation can success (figure 1.f) or can fail (figure 1.g). The SRG segmentation fails (figure 1.h) if the seed is placed outside the ROI (figure 1.d).

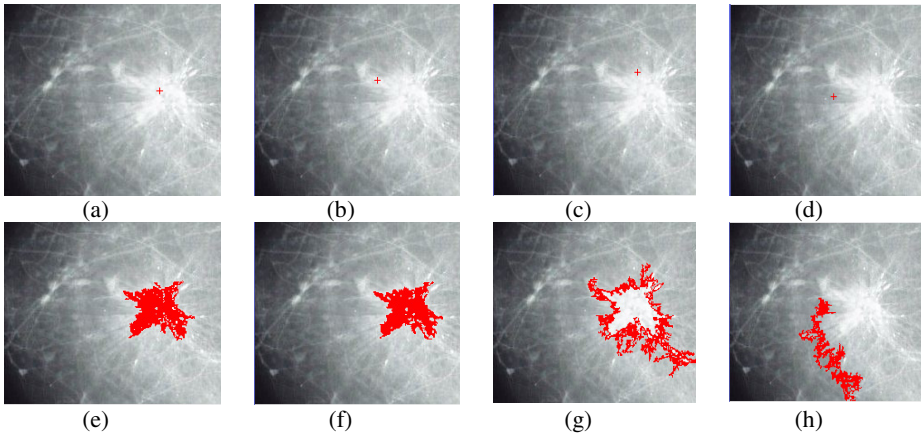


Fig. 1. Examples of different seed placement (left column) and correspondent segmentation results (right column)

To sum it up, the seed position can be considered as a good criterion in the comparison between automatic seed generation methods. The seed position is adequate if and only if the seed is placed inside the ROI. In addition, the best method is the method which generates seeds close the lesion center.

4 Results and Discussion

Considering the fact that the initial seed selection has a great influence on the final segmentation accuracy, we propose a comparative study of two automatic seed selection methods: SSFS and SSTT. The behavior of the two methods was examined using a randomly selected dataset from MiniMIAS database and Rider database. We notify that, in region growing segmentation process, the same similarity measure and the same threshold value have been used for the two methods.

4.1 Mammograms dataset

To evaluate the performance of the experimental methods, 28 mammograms with tumors are taken from MiniMIAS database. The two methods are applied on each tested image.

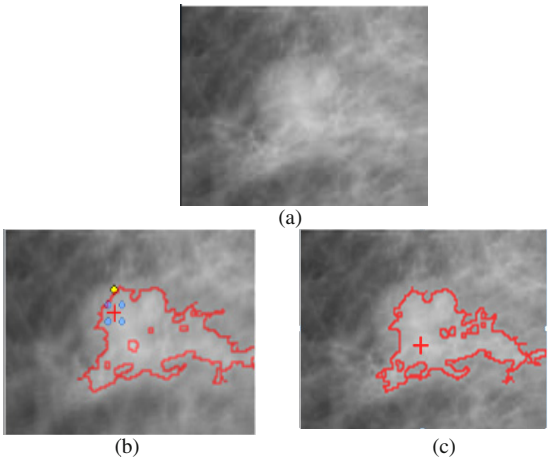


Fig. 2. Seed point generation example on mammogram. (a) Original image. (b) Seed generation result and segmentation result using SSFS. (c) Seed generation result and segmentation result using SSTT.

Figure 2 presents the results of the experimental methods on a mammogram test example. The original image is illustrated in Figure 2(a). Figure 2(b) shows seed generation result and segmentation result using SSFS method. Figure 2(c) shows seed generation result and segmentation result using SSTT method. We can see clearly that the two methods place correctly the seed inside the ROI, but at different positions.

According to the obtained results from all the tested mammograms, the seeds repartition area of each method is surrounded in figure 3. On a prototype image we have delimited separately the zones covered by SSFS and SSTT methods. The blue line delimits the SSFS zone and the red line delimits the SSTT zone. From this illustration three observations can be made:

1. The two methods SSFS and SSTT succeed in placing some seeds inside and close to the center of the ROI.
2. The SSTT method fails in some cases because it places a number of seeds outside the ROI.
3. The SSFS method gives better results than the SSTT method because in the worst case, the seed point is placed on the ROI boundary.

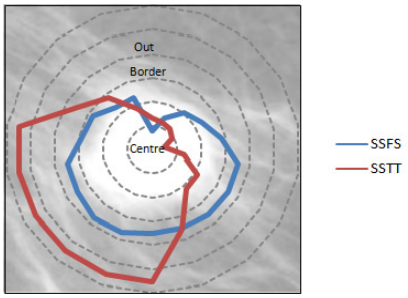


Fig. 3. Seeds repartition areas illustration for mammograms dataset

A priori, we can suppose that the SSFS method is more powerful than the SSTT method. But, when we look to the plot of the figure 4, this supposition becomes weak. The plot shows that the SSTT method places most seeds inside the ROI while the SSFS method places the majority of seeds on the ROI boundary.

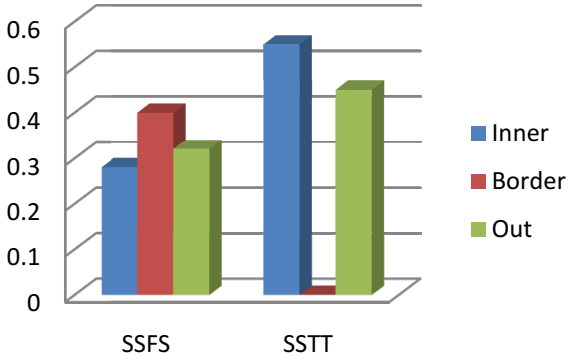
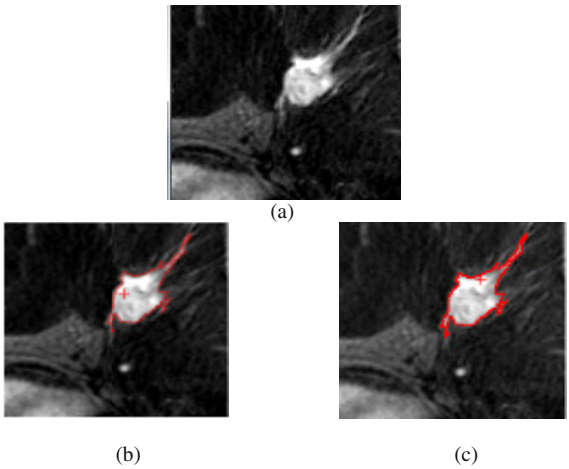


Fig. 4. Static results of the SSFS method and the SSTT method on the mammograms dataset

From the results above, we conclude that the SSFS method can easily find the ROI but, has some difficulties to point their centre. On the contrary, the SSTT method is more powerful in locating the centre area if it success in detecting the ROI.

4.2 Rider Dataset

To evaluate the performance of the experimental methods on another dataset, 20 breast IRM images with tumors were taken from Rider database. Seed point generation example by the two considered methods is shown in the following:



(a)

(b) (c)

Fig. 5. Seed point generation example on breast IRM image (a) Original image (b) Seed generation result and segmentation result using SSFS. (c) Seed generation result and segmentation result using SSTT.

Figure 5 illustrates seed generation results and segmentation results using the two methods (SSFS and SSTT) on breast IRM example. The original image is illustrated in Figure 5(a). Figure 5(b) shows seed generation result and segmentation result using the SSFS method. Figure 5(c) exhibits seed generation result and segmentation result using the SSTT method. This example shows that the SSTT method places its seed farther from the centre of ROI than the SSFS method. This fact is not correct for all the seeds generated by the SSTT method. As it is presented in figure 6, the SSTT method places all its seeds in an area (represented by red line) included in the repartition area (represented by blue line) of the SSFS method. So, the SSFS method gives better results than the SSFS method in most cases.

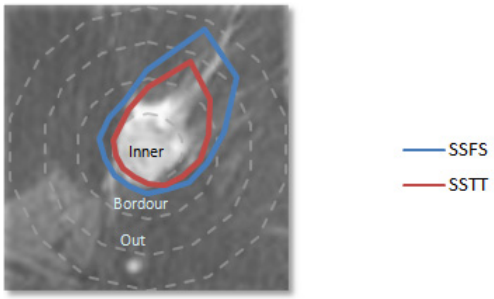


Fig. 6. Seeds repartition areas illustration for IRM dataset

The statistical data presented by the plot of the figure 7 confirms the efficiently of SSTT method in comparison with the SSFS method.

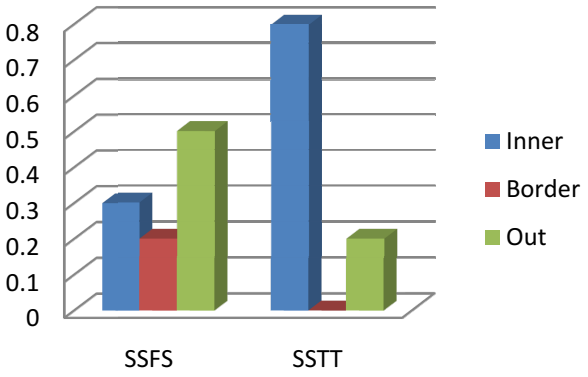


Fig. 7. Static results of the SSFS method and the SSTT method on the breast IRM dataset

4.3 Discussion

In this work, two automatic seed selection methods have been studied and evaluated. The SSFS method and the SSTT method are tested using mammograms and breast IRM images. From the obtained results some conclusions can be drawn:

1. It is possible to apply the SSFS method and the SSTT method for both modalities. The SSTT method introduced for the IRM breast images gives good results for the mammograms as well. Despite the SSFS method had been developed, originally, for mammograms it gave also acceptable results for IRM breast images.
2. The SSTT method performs well if there are no undesirable regions. Undesirable regions are the regions with high intensity like labels, artifacts ...etc. So, if these undesirable regions are removed by using pre-processing stage, the performance of the SSTT method will certainly increase.
3. Masses predefined features values were carefully studied by the authors of the SSFS method. These references values which allow good tumors detection in mammograms can be inappropriate for IRM images. Hence, references values must be modified for each new used database this repeated modification will be an obstacle for the SSFS method adaptability. However, if references values are carefully selected, the SSFS performance will consequently augment. Unfortunately, it is very hard to fix the best references values for each used database.
4. The SSTT method has proved to be more efficient in matter of spotting the ROI centre compared to the SSFS method. The SSTT method selects the high intensity pixel as a seed, while the SSFS method selects the first pixel of the mask as a seed. The SSTT method seed selection criterion makes it possible to place the seed close to the ROI centre. On the other hand, the SSFS method seed selection criterion favours the seed placement on the ROI boundary.

5 Conclusion

Since region growing technique often gives good segmentation results that correspond well to the observed edges, it is widely used in medical images. Typically, a seeded region growing algorithm includes two major steps. The first step is seed point generation by selecting an initial seed point somewhere inside the suspected lesion. The second step is region formation which starts from the seed point and grows progressively to fill a coherent region. As, region growing results are sensitive to the initial seeds, the accurate seed selection is very important for image segmentation. In this work, we have implemented, tested and evaluated two automatic seed selection methods. The SSTT method proposed by Al-Faris et al. is based on the thresholding technique. The SSFS method proposed by Yuvaria et al. is based on features similarity. The tests were elaborated on two different kinds of breast images modalities: mammograms and IRM. Both the SSTT and the SSFS methods deal well with mammograms. But, as far as IRM is concerned, the SSTT method performs better than SSFS method.

References

1. Jianping, F., Guihua, Z., Body, M., Hacid, M.S.: Seeded region growing: an extensive and comparative study. Pattern Recognition Letters 26(8), 1139–1156 (2005)
2. Deboeverie, F., Veelaert, P., Philips, W.: Image segmentation with adaptive region growing based on a polynomial surface model. Journal of Electronic Imaging 22(4), 1–13 (2013)
3. Najarian, K., Splinter, R.: Biomedical signal and image processing, 2nd edn. CRC Press, Taylor & Francis Group, United States of America (2012)
4. Massich, J., Meriaudeau, F., Pérez, E., Marti, R., Oliver, A., Marti, J.: Seed selection criteria for breast lesion segmentation in Ultra-Sound images. In: Workshop on Breast Image Analysis in Conjunction with MICCAI, pp. 57–64 (2011)
5. Mehnert, A., Jackway, P.: An improved seeded region growing algorithm. Pattern Recognition Letters 18(10), 1065–1071 (1997)
6. Jianping, F., Yau, D.K.Y., Elmagarmid, A.K., Aref, W.G.: Automatic image segmentation by integrating color-based extraction and seeded region growing. IEEE Trans. Image Process. 10(10), 1454–1466 (2001)
7. Al-Faris, A.Q., Umi Kalthum, N., MatIsa, N.A., Shuaib, I.L.: Computer-Aided Segmentation System for Breast MRI Tumour using Modified Automatic Seeded Region Growing (BMRI-MASRG). J. Digit. Imaging 27, 133–144 (2014)
8. Yuvarai, K., Ragupathy, U.S.: Automatic Mammographic Mass Segmentation based on Region Growing Technique. In: 3rd International Conference on Electronics, Biomedical Engineering and its Applications (ICEBEA 2013), Singapore, pp. 29–30 (April 1, 2013)
9. Mesanovic, N., Huseinagic, H., Kamenjakovic, S.: Automatic Region Based Segmentation and Analysis of Lung Volumes from CT Images. International Journal of Computer Science and Technology 4(2), 48–51 (2013)

10. US National Cancer Institute: reference image database to evaluate therapy response (RIDER) MRI breast, 2007 The Cancer Imaging Archive (TCIA), http://cancerimagingarchive.net./about-archive.html
11. http://peipa.essex.ac.uk/info/mias.html
12. Mohd Saad, N., Abu-Bakar, S.A.R., Muda, S., Mokji, M., Abdullah. A.R.: Automated Region Growing for Segmentation of Brain Lesion in Diffusion-weighted MRI. In: Proceeding of the International MultiConference of Enginneers and Computer Scientists, Hong Kong, vol. 1, pp. 14–16 (March 2012)

IHBA: An Improved Homogeneity-Based Algorithm for Data Classification

Fatima Bekaddour(✉) and Chikh Mohammed Amine

Abou Bekr Belkaid University, Tlemcen , Algeria
fatima.bekaddour@gmail.com,
am_chikh@yahoo.fr

Abstract. The standard Homogeneity-Based (SHB) optimization algorithm is a metaheuristic which is proposed based on a simultaneously balance between fitting and generalization of a given classification system. However, the SHB algorithm does not penalize the structure of a classification model. This is due to the way SHB's objective function is defined. Also, SHB algorithm uses only genetic algorithm to tune its parameters. This may reduce SHB's freedom degree. In this paper we have proposed an Improved Homogeneity-Based Algorithm (IHBA) which adopts computational complexity of the used data mining approach. Additionally, we employs several metaheuristics to optimally find SHB's parameters values. In order to prove the feasibility of the proposed approach, we conducted a computational study on some benchmarks datasets obtained from UCI repository. Experimental results confirm the theoretical analysis and show the effectiveness of the proposed IHBA method.

Keywords: Metaheuristics · HBA · Improvement · Machine Learning · Medical Informatics

1 Introduction

Nowadays, metaheuristics approaches represent a well-established method toward solving complex and challenging optimization problems. Offering suboptimal (optimal) quality solutions in a reasonable time, they may be considered as complement to exact optimization methods. Among popular metaheuristics, there are: Genetic Algorithm [1] emulates Darwinian evolution theory; Simulated Annealing imitates annealing process of melts [2] and Particle swarm optimization stems from biology where a swarm coordinates itself in order to achieve a goal [3].

Recently, Pham and Triantaphyllou [4][5][6]developed a new metaheuristic called HBA: Homogeneity-Based Algorithm. The Standard HBA metaheuristic (SHB) is used in conjunction with traditional data mining approaches (such as: ANN: Artificial Neural Network, DT: Decision Tree…) .The main idea of SHB algorithm is to simultaneously balance both fitting and generalization [5] by adjusting classification model through the use of the concept of Homogenous Set and Homogeneity Degree [4].This is done in order to reduce the total misclassification cost of the inferred models. However, a problem with SHB algorithm is that may not adopt computational complexity

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 129–140, 2015.
DOI: 10.1007/978-3-319-19578-0_11

of the used classification model. This is due to the way objective function is defined. For the SHB metaheuristic, the total misclassification cost is described by computing only the three type of errors (false positive, false negative and the unclassifiable cases) with their penalty costs. Additionally, for this metaheuristic, only Genetic Algorithm (GA) is adopted to find optimally thresholds values, used to control the balance between the fitting and the generalization. This may reduce SHB's freedom degree.

In this article, we extend works in [4][5][6]. New contributions lies in (1) modifying the SHB's objective function to support structural complexity of the used classifier model (2) Proposing a meta-optimization based solution to the problem of tuning SHB's parameters. The IHBA (Improved Homogeneity-Based) algorithm enhances average results obtained in comparison to the standalone algorithms. Rest of this paper is organized as follows:

The standard HBA metaheuristic (SHB) is presented in the following section, before the proposed approach IHBA is elaborated. Section 3 describes some famous benchmark datasets used to test the proposed approach and explains respective results. Last section concludes the paper.

2 Methodology

2.1 Standard Homogeneity Based-Algorithm (SHB)

SHB is a recent metaheuristic, developed by Pham and Triantaphyllou in [4][5][6]. The main idea of SHB algorithm is to adopt a simultaneously balance between generalization in order to minimize total misclassification cost (TC) [4][5][6]. Let C_{FP}, C_{FN}, C_{Uc} be the penalty costs for the false positive, false negative and unclassifiable cases respectively. Also, let us denote RateFP, RateFN, RateUc as the false positive, false negative, unclassifiable rates, respectively. Then TC is defined as follow:

$$TC=\min (C_{FP}*Rate_{FP}+C_{FN}*Rate_{FN}+C_{UC}*Rate_{UC}) \quad (1)$$

SHB algorithm is used in conjunction with data mining techniques to create classification system that would be optimal in term of TC value. There is a fundamental key issue regarding the SHB algorithm [4][5][6]:

- The more compact and homogenous decision regions are, the more accurate the inferred models are. In addition, the denser the decision regions are, the more accurate the inferred models are.

The density measurement for a homogenous set is called Homogeneity Degree (HD) [4]. In [4][5][6] ,the authors proposed a way to compute HD as follow:

$$HD= \ln (nc) / h \quad (2)$$

Where *nc* is the number of points in a given set **C**, and *h* is defined in **Heuristic rule.**

The SHB algorithm stops when all of the homogenous sets have been treated. Note that SHB metaheuristic utilize GA (Genetic Algorithm) to find optimal values of the controlling threshold: β^-, β^+, α^-, α^+.

Heuristic Rule: if ***h*** is set equal to the minimum value in set ***C*** and this value is used to compute the density ***d(x)*** using equation 3, then ***d(x)*** approaches to a true density.

$$d(x) \approx \frac{1}{n * h^D} \sum_{i=1}^{n} \prod_{m=1}^{D} \varphi\left(\frac{x^m - x_i^m}{h}\right) \quad (3)$$

Where φ is the kernel function, defined in D-dimensional space and n is the number of points in a given set **C.**
The following pseudo-code describes the SHB algorithm:

```
Start
Initial parameters setting (α+, α-, β+, β -).
1. Apply a Data Mining approach on a training dataset T1
to infer positive and negative classification models.
2. Break the inferred models into hyper spheres.
3. For each hyper sphere C do:
  Determine whether C is homogenous or not.
     If so, computer HD using formula 2.
     Else fragment C into smaller hyper spheres.
4. Sort HD in decreasing order.
5. For each homogenous set C do:
  If [(HD ≥ β+(β-)] then
     Expand C using HD and α+(α-).
  Else
     Break C into smaller homogenous sets.
end
```

2.2 A Modified SHB Objective Function

As presented above, SHB algorithm modifies an existing classification pattern such that the total misclassification cost TC (formula 1), will be optimized or significantly reduced. Nevertheless, SHB metaheuristic objective function neglects the structural complexity of a given classification model. For example, The ANN (Artificial Neural Network) structural complexity is defined as the total number of weights and bias, figured in its architecture and the time needed for network learning. It is proved by choosing theses parameters effectively minimize the network error and perform better results.

In this regards, we have proposed a modified objective function, adopting the computational complexity design function [7] to compute the penalty of a given pattern classification architecture as follow:

$$fobj = Penality * \frac{\alpha 1 * TC_{Training} + \alpha 2 * TC_{Generalization}}{\alpha 1 + \alpha 2} \quad (4)$$

Where $(\alpha 1, \alpha 2) > 0 \in \Re$ (usually $\alpha 1 \leq \alpha 2$), are factors indicating importance degree of the learning and the generalization errors respectively. Penalty presents the model architecture influence of the objective function value as follow [7]:

$$Penality = 5 * \ 10^{-8} * e^{f(x)} + 5 * 10^{-5} * y + 1 \quad (5)$$

Where: y is the number of epochs necessary in the model training; f(x) is the Structural complexity of a classification model.
Using different values of C_{FP}, C_{FN}, C_{Uc} in objective function formula (4) , we design others objective functions formula (6-7-8) as follows:

$$fobj = Penality * \frac{\alpha1*(RateFP_{Train}+RateFN_{Train})+\alpha2*(RateFP_{Gener}+RateFN_{Gener})}{\alpha1+\alpha2} \quad (6)$$

$$fobj = Penality * \frac{\alpha1*TC1_{Train}+\alpha2*TC1_{Gener}}{\alpha1+\alpha2} \quad (7)$$

Where: $TC1_{Train} = 3RateFP_{Train} + 3RateFN_{Train} + 3RateUc_{Train}$

$$TC1_{Gener} = 3RateFP_{Gener} + 3RateFN_{Gener} + 3RateUc_{Gener}$$

$$fobj = Penality * \frac{\alpha1*TC2_{Train}+\alpha2*TC2_{Gener}}{\alpha1+\alpha2} \quad (8)$$

Where: $TC2_{Train} = RateFP_{Train} + 20RateFN_{Train} + 3RateUc_{Train}$

$$TC2_{Gener} = RateFP_{Gener} + 20RateFN_{Gener} + 3RateUc_{Gener}$$

Note that, ($RateFP_{Train}$, $RateFN_{Train}$, $RateUc_{Train}$) represent FP, FN and Uc rates during the training phase and ($RateFP_{Gener}$, $RateFN_{Gener}$, $RateUc_{Gener}$) represent FP, FN and UC rates during the test phase.

- **In Formula 6:** we do not penalize Uc, but penalize the same cost for FP, FN.
- **In Formula 7:** we penalize all three error types by unit equal to three.
- **In Formula 8:** we penalize more FN than the other type of errors.

2.3 Tuning SHB Parameters by Means of Metaheuristics

Within the scope of SHB algorithm, there are four parameters which are used to control the balance of fitting and generalization that would minimize (or significantly reduce) the total misclassification cost (TC):

- Two expansion factors α^-, α^+, to be used for expanding the negative and the positive homogenous sets.
- Two breaking factors β^-, β^+, to be used for breaking the negative and the positive homogenous sets.

Note that, if the expansion parameters values (α^-, α^+) are too high, then this would result in the oversimplification problem. On the contrary, too low expansion parameters values may not be sufficient to overcome the overfitting problem. The opposite situation is true with the breaking factors values (β^-, β^+). Authors in [4][5][6] propose to only use genetic algorithm(GA) to find optimal threshold values for α^-, α^+, β^-, β^+. This may reduce the freedom degree of the SHB algorithm .

This article employs several metaheuristics approaches to formally test the existence of a relationship between performance and effective parameters values. In par-

ticular, (PSO: Particle Swarm Optimization, SA: Simulated Annealing and GA: Genetic Algorithm) metaheuristics are used for the SHB algorithm parameters α-, α+, β -, β+. That is these parameters represents individual variables and *fobj* described in formula 4 is taken as objective function. Since PSO, SA and GA metaheuristics approaches are tested using a dataset to find optimal values for (α-, α+, β -, β+), a calibration dataset is needed. This requirement can be fulfilled in the following way: the original training dataset T is divided into two datasets: T1 (for example: 90%) for training data mining models to infer positive and negative classification models, and T2 as a calibration dataset.

In the first phase, hyperspheres that cover decision regions are employed to obtain homogenous set (using step 3to 5 described in the pseudo-code of SHB algorithm) . Then, classification models (homogenous sets) are evaluated by using the calibration dataset T2 to compute *fobj*. Next, metaheuristic bloc could replace the default tuning parameters GA (Genetic Algorithm) and determine the new threshold values (α^-, α^+, β^-, β^+).

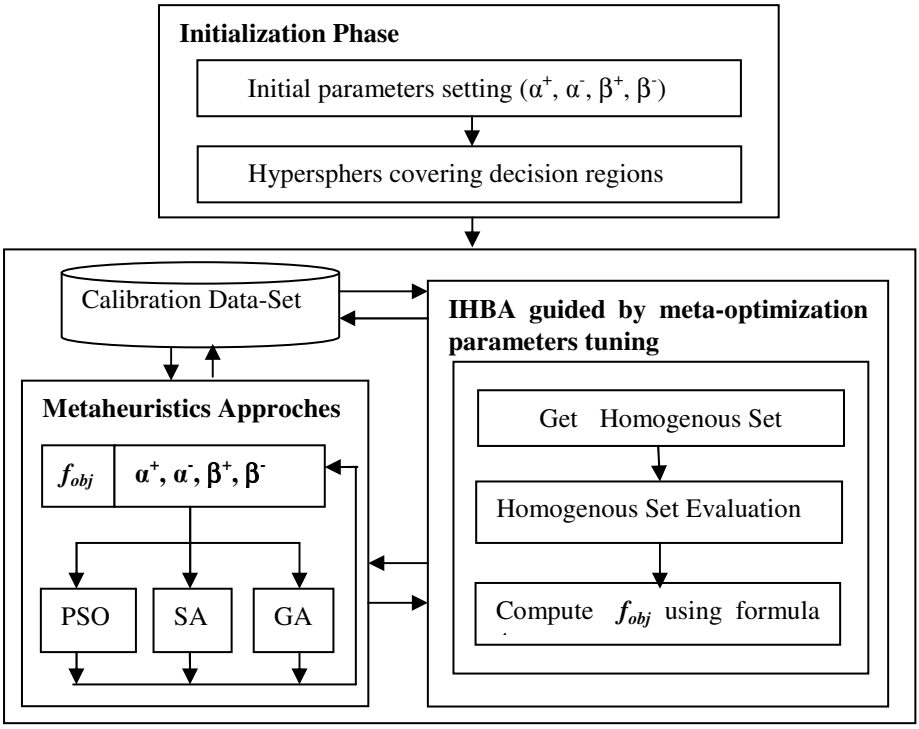


Fig. 1. Architecture of the proposed system to determine IHBA parameters

In fact, this leads to a meta-optimization approach, which means that any metaheuristic is used to search for the best tuning of parameters of metaheuristic in solving a given optimization problem [3]. After a number of iterations, the proposed approach returns the optimal threshold values of (α^-, α^+, β^-, β^+). It is to be emphasized that by

employing metaheuristics bloc during SHB algorithm iterations , permit to estimate effective parameters setting for SHB metaheuristic and therefore, allow to approximate a functional relationship between classifier's performance and effective parameters . The architecture of the overall system is depicted in figure 1.

3 Some Computational Results

3.1 Benchmark Data Sets

This paper studies two medical data sets: Appendicitis (AP), and Thyroid (TR). **Table 1** shows a summary of the main characteristics of these datasets. The benchmark chosen present a variety of descriptions (including number and type of attributes, number of instances…). The first dataset is Appendicitis, created by Kapouleas and Weiss (1989) [8] from Rutgers university. The features were obtained from laboratory tests as follow: WBC1, MNEP, MNEA, MBAP, HNEP and HNEA. The second medical dataset is thyroid disease, obtained from UCI repository [9]. It consists of five continuous attributes. The task is to identify whether a patient is normal or suffers from hyper (hypo) -thyroidism.

Table 1. Medical datasets characteristics

Datasets	No. Instances	No. Features	Training Dataset	Testing Dataset
Appendicitis	106	7	79	27
Thyroid	215	5	143	72

3.2 Results and Discussion

The following are some computational results obtained from several experiments performed for each data mining approaches used in such work. Experiments were conducted with two datasets obtained from UCI repository [9]. As discussed before, we considered three scenarios for the IHBA objective function. Also, we choose different setting for (α_1, $\alpha 2$) factors that are used to weigh the importance degree attributed to the learning and the generalization errors respectively.

Initially, we assigned an equal weight ($\alpha_1=\alpha_2=1$) to the learning and the generalization errors for ANFIS (Adaptative Neuro-Fuzzy Inference System) [10], LVQ (Learning Vector Quantization) [11] and PMC (Perception Multi-layers). Then, we choosed a larger weight to the generalization than the training error ($\alpha_1=0.5$; $\alpha_2=1$). Finally, we attributed more importance to the ability of learning than the ability of finding correct output value for an unknown data sample ($\alpha_1=1$; $\alpha_2=0.5$).The results of these simulations are shown in Table 2, 3 and 4. According to those tables, it appears that α_1 and $\alpha 2$ factors have influence on the final results. Those Tables show the misclassification testing error rate ($\boldsymbol{TC_{Test}}$) and $\boldsymbol{f_{eval}}$ (the objective function evaluation) obtained for original algorithms (ANFIS, LVQ, PMC) and the proposed IHBA

approach. The colon improvement presents any improvement rate achieved by the IHBA when compared with that of the standalone algorithm.

Table 2. Results in minimizing (f_{eval}=*FP+ FN)*

Datasets	Alg	α1 α2	Original-Alg TC_{Test}	Original-Alg f_{eval}	IHBA TC_{Test}	IHBA f_{eval}	Improv (%)
AP	ANFIS	1 1	3	26.9	11.1	30.9	No.impr
		0.5 1	3	19.3	3.7	19.8	No.impr
		1 0.5	3	38.2	0.00	37.1	2.87
	LVQ	**1 1**	11	14.4	3.7	10.7	25.69
		0.5 1	3	7.08	3.7	7.55	No.impr
		1 0.5	11	13.8	3.7	11.3	18.11
	PMC	1 1	7	16.3	7.4	16.5	No.impr
		0.5 1	3	7.86	7.4	10.9	No.impr
		1 0.5	14	18.2	7.4	15.8	13.18
TR	ANFIS	1 1	69	40.8	30.5	21.5	47.30
		0.5 1	69	51.1	27.7	23.6	53.81
		1 0.5	69	35.1	27.7	21.3	39.31
	LVQ	1 1	25	23.4	26.3	24.0	No.impr
		0.5 1	69	69.6	26.3	41.0	41.09
		1 0.5	25	22.8	26.3	23.2	No.impr
	PMC	1 1	4	2.78	5.5	3.6	No.impr
		0.5 1	2	2.1	1.3	1.61	23.33
		1 0.5	2	2.6	5.5	3.81	No.impr

In a first scenario (formula 6), we did not penalize for the unclassifiable cases (Uc), and penalized by one unit the FP (False Positive) and the FN (False Negative) errors. The results of this scenario are shown in Table 2.This table shows that the average values of f_{eval} obtained from the IHBA on the AP and TR datasets were 17.83, 18.18 respectively. Furthermore, theses values of f_{eval} were optimal than the average values of f_{eval} achieved by the stand-alone algorithms on AP and TR datasets by about 6.65, 22.76 respectively.

In the second scenario (formula 7), we assumed that all three error types would be penalized by an identical value, equal to three units. The results are presented in Table 3. The average values for f_{eval} obtained from IHBA on AP, TR datasets were 147.55, 75.75 respectively. These values for f_{eval} were less than the average values of f_{eval}

Table 3. Results in minimizing (f_{eval}= 3*FP*+ 3*FN*+3*Uc*)

Datasets	Alg	α1 α2	Original-Alg TC_{Test}	Original-Alg f_{eval}	IHBA TC_{Test}	IHBA f_{eval}	Improv (%)
AP	ANFIS	1 1	233.33	193.1	288.9	221.0	No.impro
		0.5 1	233.33	208.0	300.0	252.6	No.impro
		1 0.5	233.33	189.6	300.0	211.9	No.impro
	LVQ	**1 1**	233.33	144.2	166.6	110.6	23.3
		0.5 1	233.33	171.8	144.4	112.1	34.7
		1 0.5	233.33	108.7	144.4	78.88	27.4
	PMC	1 1	233.33	160.9	155.5	119.9	25.5
		0.5 1	233.33	181.2	155.5	126.5	30.2
		1 0.5	233.33	121.8	155.5	94.47	22.4
TR	ANFIS	1 1	220.8	129.3	91.66	64.64	50.0
		0.5 1	220.8	162.6	95.83	79.26	51.5
		1 0.5	220.8	109.9	95.83	68.21	37.9
	LVQ	1 1	283.3	174.9	83.33	74.43	57.4
		0.5 1	220.8	218.2	83.33	126.0	No.impro
		1 0.5	283.3	138.3	83.33	71.33	48.4
	PMC	1 1	212.5	113.6	108.3	58.90	48.1
		0.5 1	216.7	153.8	133.3	95.41	37.9
		1 0.5	212.5	80.04	108.3	43.57	45.5

achieved by original algorithms by about 18.16 and 41.8 on the AP, and TR datasets respectively. In the last scenario (formula 8), we assumed that the FN would be more penalized than the other two types of errors (FP, FN). In particular, table 4 shows that the average values for f_{eval} obtained from IHBA on the AP and TR datasets were 236.63, 94.29 respectively .This table, shows that the f_{eval} were less than the original algorithms (ANFIS, LVQ, PMC) by about 0.43, 63.23 when applied on the AP and TR datasets respectively.

When comparing the tables 2, 3, 4, it appears that PMC and ANFIS models, usually obtain better results. However, ANFIS is more practical due to its transparency. Additionally, in some cases, the f_{eval} value of a standalone approach yielded better values than the one achieved by the IHBA metaheuristic. A reason for that is that the standalone algorithm may have reached the global optimal value (or close to that) for f_{eval} .Note that the number of membership functions and hidden layers affect the structural complexity of the neuro-fuzzy system and the artificial neural network models respectively, in this work, we proposed to use two membership function for ANFIS system and one hidden layer for LVQ and PMC classification models .

The best architecture model found for ANFIS, LVQ and PMC models were (128,20,20) for AP and (32,20,20) for TR dataset respectively.

Table 4. Results in minimizing (f_{eval}=*FP+ 20FN+3Uc)*

Datasets	Alg	α1 α2	Original-Alg		IHBA		Improv (%)
			TC_{Test}	f_{eval}	TC_{Test}	f_{eval}	
AP	ANFIS	1 1	296.29	174.0	300.00	175.9	No.impr
		0.5 1	296.29	215.5	292.59	213.0	1.16
		1 0.5	296.29	136.3	288.88	133.9	1.76
	LVQ	1 1	211.11	187.8	281.48	223.3	No.impr
		0.5 1	225.92	229.4	281.48	266.7	No.impr
		1 0.5	211.11	194.0	281.48	217.7	No.impr
	PMC	1 1	288.88	367.6	281.48	363.8	1.03
		0.5 1	225.92	265.9	281.48	304.9	No.impr
		1 0.5	203.7	203.2	281.48	230.5	No.impr
TR	ANFIS	1 1	1327.6	670.8	30.55	21.6	96.8
		0.5 1	1327.6	891.0	30.55	25.52	97.1
		1 0.5	1327.6	455.1	40.27	25.56	94.4
	LVQ	1 1	221.05	121.9	30.55	26.24	78.4
		0.5 1	1327.6	1358	30.55	488.9	63.9
		1 0.5	221.05	88.57	30.55	24.8	71.9
	PMC	1 1	222.36	12.47	97.22	51.8	No.impr
		0.5 1	225.00	158.2	155.5	109.6	30.7
		1 0.5	221.05	116.5	101.3	74.64	35.9

In order to shed some light upon the second contribution, Table 5 provides an overview of the results obtained throughout the empirical comparison of different meta-optimization based solution (PSO, SA and GA) to the problem of tuning SHB's parameters. The colon improvement 1 shows any improvement of f_{eval} achieved by IHBA enhanced by means of metaheuristics approaches to find optimal thresholds values (α^+, α^-, β^+, β^-), when compared with the standalone algorithms under the first consideration (where α_1=0.5, α_2=1) and by using ANFIS model. The colon improvement 2 shows any improvement of f_{eval} achieved by IHBA enhanced by means of meta-optimization parameters tuning, when compared with best results obtained with IHBA under the first consideration, where α_1,=0.5; α_2=1. We have simulated this scenario (α_1=0.5, α_2=1), because it seems to be more realistic that the ability to learn the model is less relevant than the ability to generalize (i.e. find a correct output value for an unknown data sample).

The colon parameters setting specify different parameters configuration for considered metaheuristics (PSO, SA and GA). In particular, PSO algorithm has been applied with different values of number of iterations (50, 100), population size (20, 40), social attraction and cognitive attraction(0.25, 0.7). In case of SA metaheuristic, we optimized SHB's factors (α^+, α^-, β^+, β^-) by setting different values of iteration number (500, 1000) and the perturbation function. The initial temperature was set either to 50

or 100. In the GA, each chromosome encodes the two expansion thresholds values (α^+, α^-) and the two breaking thresholds values (β^+, β^-). The population evolves in search for the optimal values of these parameters. We have applied the GA with different values of: number of generation (200, 1000), population size (15, 35) and crossover fraction (0.5, 0.7). Mutation fraction equaled 0.01.

Table 5. Results of IHBA improved by means of parametes tuning (f_{eval}=*FP+ FN)*

Data sets	Meta-Heuristic	Parameters Setting	Results TC_{Test}	Results f_{eval}	Improv1 Rate	Improv2 Rate
AP	PSO	100 ; 20 ; 0.7 ; 0.25	3.7	37.1	No.impr	No.impro
		50 ; 20 ; 0.25 ; 0.7	7.4	39.6	No.impr	No.impro
		50 ; 40 ; 0.7 ; 0.25	3.7	37.1	No.Impr	No.Impr
		100 ; 40 ; 0.25 ; 0.7	0.00	34.6	No.impr	No.impro
	SA	500;50 ;Fast	3.7	37.1	No.impr	No.impro
		1000;100. Fast	3.7	37.1	No.impr	No.impro
		1000 ;50 ; Bolz	14.8	44.5	No.impr	No.impro
		500 ;100 ;Bolz	3.7	37.1	No.impr	No.impro
	GA	15; 200; 0.5; 0.01	0.00	34.6	No.impr	No.impro
		35;1000; 0.7 ;0.01	0.00	34.6	No.impr	No.impro
		35; 200; 0.5; 0.01	3.7	37.1	No.impr	No.impro
		15;1000; 0.7 ;0.01	0.00	34.6	No.impr	No.impro
TR	PSO	100 ; 20 ; 0.7 ; 0.25	2.77	12.0	76.4	48.9
		50 ; 20 ; 0.25 ; 0.7	26.4	27.8	45.6	No.impro
		50 ; 40 ; 0.7 ; 0.25	9.60	16.7	67.3	29.23
		100 ; 40 ; 0.25 ; 0.7	25.0	26.9	51.0	No.impro
	SA	500;50 ;Fast	2.77	12.0	76.4	48.9
		1000;100. Fast	2.77	12.0	76.4	48.9
		1000 ;50 ; Bolz	2.77	12.0	76.4	48.9
		500 ;100 ;Bolz	2.77	12.0	76.4	48.9
	GA	15; 200; 0.5; 0.01	12.5	18.5	63.7	21.4
		35;1000; 0.7 ;0.01	11.1	17.6	65.5	25.4
		35; 200; 0.5; 0.01	2.77	12.0	76.4	48.9
		15;1000; 0.7 ;0.01	11.1	17.6	65.5	25.4

It is clearly visible, that the PSO metaheuristic achieved better results (in minimizing f_{eval}), for big number of iterations, population size and cognitive attraction. Simulated annealing algorithm was slightly worse than GA metaheuristic.

Table 5 shows that the average values of f_{eval} obtained from IHBA approach improved by means of meta-optimization approaches on AP and TR datasets were 37.09 and 16.45 respectively. In addition, these values of f_{eval} were less than those achieved by standalone methods and IHBA approach depicted in Table 2 on TR dataset by about 68.08 and 32.9 respectively. Note that, The proposed IHBA approach

improved by means of parameters tuning based on (PSO, SA and GA) metaheuristics, when applied on AP dataset, found no improvement of f_{eval} , compared to original results depicted in Table 2. A reason for that, is that the standalone approaches or IHBA may have achieved optimal (or near-optimal) values of f_{eval}.

4 Conclusion

Considering importance of parameters tuning of a given metaheuristic algorithm, in this paper, we proposed an Improved Homogeneity Based-Algorithm which uses computational complexity of a classifier model as a modified objective function. Additionally, we employed several metaheuristics approaches (Simulated annealing, Genetic Algorithm and Particle Swarm Optimization) to find optimally thresholds values, used to refine the inferred models regions obtained by applying a classification method. The proposed method IHBA (Improved Homogeneity-Based Algorithm) tested on some benchmarks data sets from the UCI repository indicated the increased performance of the proposed algorithm in comparison with the standalone algorithms (ANFIS, LVQ and PMC). Future works will extend the SHB metaheuristic with feature subset selection aiming to reduce classification time and making HBA applicable to higher data dimensionality.

References

1. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1992, 1975) (re-issued by MIT Press)
2. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
3. Talbi, E.-G.: Metaheuristics: From Design to Implementation. Wiley (June 2009)
4. Pham, H.N.A., Triantaphyllou, E.: The impact of overfitting and overgeneralization on the classification accuracy in data mining. In: Maimon, O., Rokach, L. (eds.) Soft Computing for Knowledge Discovery and Data Mining, part 4, ch. 5, pp. 391–431. Springer, New York (2007)
5. Pham, H.N.A., Triantaphyllou, E.: Prediction of diabetes by employing a new data mining approach which balances fitting and generalization. In: Lee, R., Kim, H.-K. (eds.) Computer and Information Science. SCI, vol. 131, pp. 11–26. Springer, Heidelberg (2008)
6. Pham, H.N.A., Triantaphyllou, E.: An application of a new meta-heuristic for optimizing the classification accuracy when analyzing some medical datasets. Expert Systems with Applications 36(5), 9240–9249 (2009)
7. Carvalho, A.R., Ramos, F.M., Chaves, A.A.: Metaheuristics for the feedforward artificial neural network (ANN) architecture optimization problem. Neural Computing and Applications 20(8), 1273–1284 (2011)
8. Weiss, S.M., Kapouleas, I.: An empirical comparison of pattern recognition, neural nets and machine learning classification methods. In: Shavlik, J.W., Dietterich, T.G. (eds.) Readings in Machine Learning. Morgan Kauffman Publ., CA (1990)

9. UCI repository of machine learning databases, University of California at Irvine, Department of Computer Science, http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease (last accessed 2015)
10. Jang, J.S.R.: Anfis: adaptative network-based fuzzy inference système. IEEE Trans. on Systems, Man and Cybernetics (1993)
11. Kohonen, T.: The Self-Organizing Map. Proceedings of the IEEE 78(9), 1464–1480 (1990)

Multiple Guide Trees in a Tabu Search Algorithm for the Multiple Sequence Alignment Problem

Tahar Mehenni(✉)

Computer Science Department,
University Mohamed Boudiaf of M'sila, 28000 M'sila, Algeria
tmehenni@univ-msila.dz

Abstract. Nowadays, Multiple Sequence Alignment (MSA) approaches do not always provide consistent solutions. In fact, alignments become increasingly difficult when treating low similarity sequences. Tabu Search is a very useful meta-heuristic approach in solving optimization problems. For the alignment of multiple sequences, which is a NP-hard problem, we apply a tabu search algorithm improved by several neighborhood generation techniques using guide trees. The algorithm is tested with the BAliBASE benchmarking database, and experiments showed encouraging results compared to the algorithms studied in this paper.

Keywords: Multiple sequence alignment · Tabu search · Neighborhood · Guide tree

1 Introduction

Multiple sequence alignment (MSA) is a very interesting problem in molecular biology and bioinformatics. Although the most important regions of DNA are usually conserved to ensure survival, slight changes or mutations (indels) do occur as sequences evolve. Methods such as sequence alignment are used to detect and quantify similarities between different DNA and protein sequences that may have evolved from a common ancestor.

Sequence alignment is the way of inserting dashes into sequences in order to minimize (or maximize) a specified scoring function [1,26]. There are two classes of sequencing; pairwise sequence alignment (PwSA) and multiple sequence alignment (MSA). The latter is simply an extension of pairwise alignments that align 3 or more sequences. Both MSA and PwSA can further be categorized as global or local methods. As global methods attempt to align entire sequences, local methods only align certain regions of similarity.

The majority of multiple sequence alignment heuristics is now handled using progressive approach [13]. Progressive also known as hierarchical or tree methods, generate a multiple sequence alignment by first aligning the most similar sequences and then adding successively less related sequences or groups to the alignment until the entire query set has been incorporated into the solution. Sequence relatedness is described by the initial tree that is based on Pair

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 141–152, 2015.
DOI: 10.1007/978-3-319-19578-0_12

wise alignments which may include heuristic Pair wise alignment methods. Some well-known programs using progressive strategies are ClustalW [28], Muscle [6], MULTAL [12] and T-COFFEE [20]. This approach has the advantages of speed and simplicity. However, its main disadvantage is the local minimum problem, which comes from the greedy nature of the approach.

Another approach is to prune the search space of the Dynamic Programming (DP) algorithm for simultaneously aligning multiple sequences, e.g., MSA [11, 18], OMA [23] etc. Algorithms of this approach often find better quality solutions than those of the progressive approach. However, they have the drawbacks of complexity, running time and memory requirement, so they can only be applied to problems with a limited number of sequences (about 10).

The iteration-based approach is also applied to the multiple sequence alignment. Iterative alignment methods produce alignment and refine it through a series of cycles (iterations) until no further improvements can be made. It is deterministic or stochastic depending on the strategy used to improve the alignment. This approach includes iterative refinement algorithms, e.g,. PRRP [10], simulated annealing [14], genetic algorithms (SAGA [19], MAGA [29]), Ant Colony [3] and Swarm Intelligence [15]. Therefore, they can evade being trapped in local minima.

In this paper, we present an iteration-based approach using tabu search features to find the global alignment of multiple sequences, where the neighbors are generated using a set of operations on the guide tree of the initial solution.

The remaining of the paper is organized as follows. In section 2, we present the related work in MSA using tabu search. Section 3, describes our algorithm. Experimental results are presented in section 4 and the study is concluded in section 5.

2 Related Work

Tabu Search (TS) [8, 9] was developed by Fred Glover in 1988. It was initiated as an alternative local search algorithm addressing combinatorial optimization problems in many fields like scheduling, computer channel balancing, cluster analysis, space planning etc. Tabu search is an iterative heuristic approach that uses adaptive memory features to align multiple sequences. The adaptive memory feature, a tabu list, helps the search process to avoid local optimal solutions and explores the solution space in an efficient manner.

In [24], authors propose a tabu search algorithm for multiple sequence alignment. The algorithm implements the adaptive memory features typical of tabu searches to align multiple sequences. Both aligned and unaligned initial solutions are used as starting points for this algorithm. Aligned initial solutions are generated using Feng and Doolittles progressive alignment algorithm [7]. Unaligned initial solutions are formed by inserting a fixed number of gaps into sequences at regular intervals. The quality of an alignment is measured by the COFFEE objective function [21]. In order to move from one solution to another, the algorithm moves gaps around within a single sequence and performs block moves.

This tabu search uses a recency-based memory structure. Thus, after gaps are moved, the tabu list is updated to avoid cycling and getting trapped in a local solution.

[17] develops in his thesis several tabu searches that progressively align sequences. He begins by a simple tabu, called Tabu A, using Dynamic Programming (DP). Then, he proposes other modified versions of tabu search, using at each time a new feature for the previous algorithm, like subgroups alignment, intensification and diversification.

In this paper, we develop a novel tabu search algorithm, by adapting similar procedures of Tabu search developed by [17], and adding a new and efficient technique for generating neighbors using guide trees.

3 Algorithm Overview

We first give a general description of the tabu search components of our method (initial solution, neighborhood generation and intensification method), and then provide a summarizing pseudo-code description of the main algorithm.

Tabu search works by starting from an initial solution, and iteratively explores the neighborhood of current solution by generating the moves called neighbors. In each iteration, the neighbors are evaluated through the alignment score and the best neighbor, provided it is not in the tabu list, is selected and applied to the current solution. This produces a new current solution for the next iteration. The applied neighbor is added to the tabu list and it is not allowed for a specified number of iteration called tabu tenure.

3.1 Initial Solution

The generation of an initial solution is an important step towards getting a final improved alignment. A good initial solution can effectively converge faster and hence cut the computational cost. The initial solution of the tabu search is represented by a tree that is generated using the neighbor-joining guide tree (NJ) [25], which fixes the order of the partial alignments in the progressive alignment.

The NJ method constructs guide trees by clustering the nearby sequences in a stepwise manner. In each step of the sequence clustering, it minimizes the sum of branch lengths, selecting the two nearest sequences/nodes and joining them. Next, the distance between the new node and the remaining ones is recalculated. This process is repeated until all sequences are joined to the root of the guide tree. Figure 1 gives an example of a guide tree produced by 5 sequences.

The MSA is obtained from the tree as follows: the pair of sequences on the lowest level are aligned first. Then, the entire branch containing these two sequences is aligned starting from the lowest level and progressing upward to sequences on higher levels. After the MSA is determined, the alignment is scored.

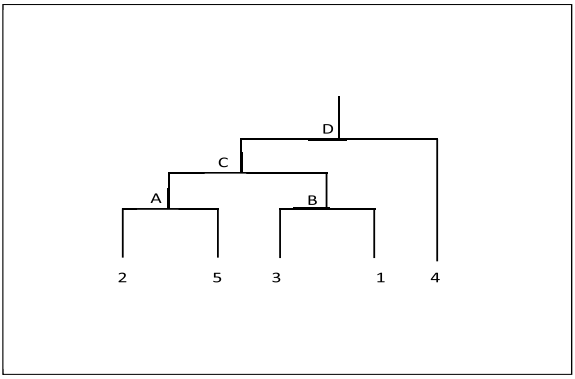


Fig. 1. An example of a guide tree generated by NJ Clustering Algorithm as Initial solution for the Tabu Search

The most popular scoring scheme is the sum of all pairwise alignments score: Sum-of-Pairs Score (SP).

$$SP = \sum_{i=1}^{n-1} \sum_{j=i}^{n} Score(S_i, S_j) \tag{1}$$

where

$$Score(S_i, S_j) = \max \begin{cases} (S_{i-1}, S_{j-1}) + s(x_i, y_j) \\ (S_{i-1}, S_j) - d \\ (S_{i-1}, S_j) + d \end{cases}$$

where $s(x_i, y_j)$ is the score for matching symbols x_i and y_j and d is the penalty for introducing a gap.

3.2 Neighborhood Generation

The neighborhood of the current solution may be generated by one of the four ways: swapping, node insertion, branch insertion or distance variation.

Generation by Swapping. The simplest way of generating a neighborhood is swapping the order of the sequences (i.e. leaves) while maintaining the same guide tree topology. the number of guide trees generated by swapping is $n(n-1)/2$, where n is the number of sequences to be aligned. Figure 2 shows two guide trees (b and c) generated from the initial guide tree a by swapping the order of the sequences.

Generation by Node Insertion. Neighbors can be generated from the current solution (i.e. the current guide tree) by performing certain insertions of nodes. The node insertion makes it possible to move a sequence node to another location

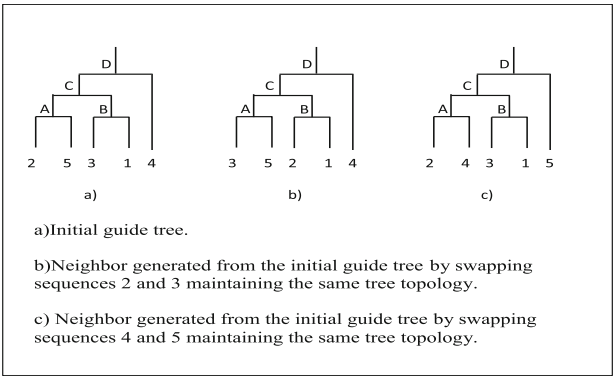


Fig. 2. Two examples of neighbors generated by swapping technique from the initial solution

of the guide tree. This will change the topology of the initial guide tree, and the new guide tree can be considered as a neighbor of the original one.

The neighborhood can be generated randomly by this technique, since the topology of the initial guide tree is not predetermined. However, we can make only n node insertions to obtain exactly n neighbors, by selecting randomly a node to share one of the sequences (leaves) of the guide tree. More precisely, for each sequence, we choose randomly a node and move it to share this sequence, and so on. Figure 3 shows two guide trees (b and c) obtained by inserting nodes to share predetermined sequences of the initial guide tree a.

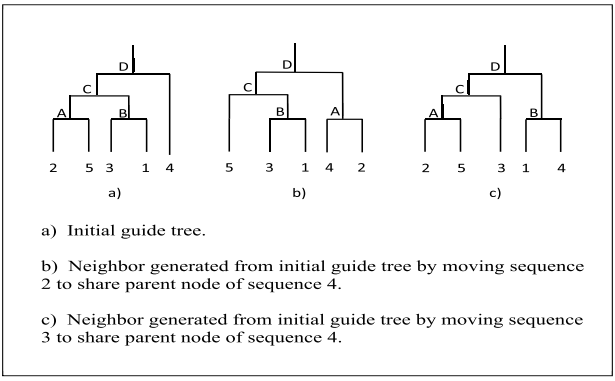


Fig. 3. Two examples of neighbors generated by node insertion technique from the initial solution

Generation by Branch Insertion. Another way to generate neighbors from the current guide tree is the branch insertion, which is moving a branch of the guide tree (or a sub-tree) to another location. The new guide tree resulting of

this move is considered as a neighbor of the current guide tree. This will change the topology of the initial guide tree.

Neighbors are generated randomly by branch insertion move. However, we can make only n branch insertions to generate exactly n neighbors for the current guide tree. For each sequence, we choose randomly a branch (or sub-tree) and move it to share this sequence, and so on. Figure 4 shows two guide trees (b and c) obtained by inserting branches to share predetermined sequences of the initial guide tree a.

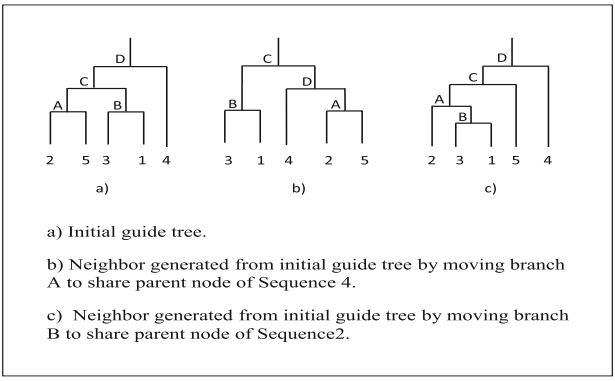


Fig. 4. Two examples of neighbors generated by branch insertion technique from the initial solution

Generation by Distance Variation. The last technique used to generate a neighborhood is the distance variation. Since the initial guide tree is obtained using NJ clustering algorithm, we can produce N different guide trees based on the NJ clustering algorithm, N being defined by the user. Each tree corresponds to a variation of the original obtained by NJ but adding some random noise into the distances in order to introduce some variability. The variation introduced in the guide tree is low enough to keep the distance criteria but significant enough to provide the necessary flexibility to generate multiple alternative trees [22]. Figure 5 shows two guide trees (b and c) produced by adding variation to distances in the NJ clustering algorithm used to obtain the initial guide tree a.

3.3 Intensification Method

Generally, an intensification procedure revisits and examines good solutions. It maintains the good portions of this solution and searches to find a better neighboring solution.

When a single MSA continues to have the highest score for many iterations, the intensification phase aims to escape the local minima by taking out a solution from the tabu list and restart another search process.

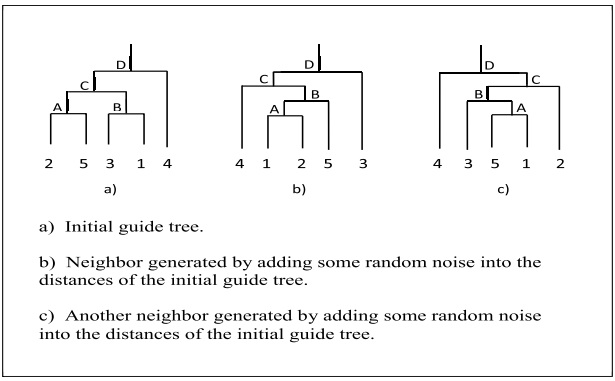


Fig. 5. Examples of neighbors generated by distance variation technique from the initial solution

3.4 Tabu Search Algorithm

Our Tabu Search algorithm consists of generating a neighborhood of a multiple sequence σ using the techniques cited above, i.e. Swapping (SWP), Node insertion (NI), Branch insertion (BI) and Distance variation (DV). The best MSA σ' having the higher score S_{max} is selected for the next iteration and put in the tabu list $TabuList$. This process is iterated until a T_{max} global running time is met . The pseudo-code of our tabu search algorithm is given in Algorithm 1. The details of this algorithm are explained below.

Algorithm 1. Tabu Search Algorithm for MSA

```
1: procedure GTreeTabu
2:     Generate σ an initial MSA using NJ algorithm;
3:     S_max:=Score(σ); σ_max := σ; TabuList:=[];
4:     while not T_max do
5:         Generate a neighborhood N(σ) using: SWP, NI, BI or DV.
6:         set σ' such that
7:         S_σ':=max_{η∈N(σ)}Score(η) and σ' ∉ TabuList
8:         if S_σ' > S_max then
9:             S_max := S_σ'; σ_max := σ'
10:            Insert σ' in TabuList
11:        end if
12:        set σ:=σ'
13:    end while
14: end procedure
```

After generating an initial solution using NJ clustering algorithm, its score is computed. While a time execution T_{max} is not reached, the tabu search is iteratively executed. Each iteration begins by generating the neighborhood of

the current solution by one of the techniques among: Swapping, Node insertion, Branch insertion, Distance variation. For each neighbor, we compute its score in order to set the best neighbor having the highest score as the new current solution. This new solution is inserted in the tabu list which has a variable length depending on the number of iterations with or without improvement. If there is improvement in a certain number of continuously iterations, the length is increased in order to insert other possible solutions. The length of tabu list is decreased if within many iterations there is no improvement. In this case, a solution will be get out from the tabu list in order to restart another search process in the intensification mode.

4 Experimental Results and Discussion

The proposed approach is implemented in MATLAB and tested on Intel Core i3-380M Laptop with 2 GB. To demonstrate the effectiveness of our approach, we have evaluated it on BAliBASE 2 benchmark base [2]. BAliBASE is a database of manually refined multiple sequence alignments. It can be viewed at http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/index.html or can be downloaded from ftp://ftp-igbmc.u-strasbg.fr/pub/BAliBASE2/.

BAliBASE database is divided into five reference sets. Reference 1 contains alignments of equidistant sequences of similar length, with no large insertions or extensions. Reference 2 aligns up to three "orphan" sequences (less than 25% identical) from reference 1 with a family of at least 15 closely related sequences. Reference 3 consists of up to 4 sub-groups, with less than 25% residue identity between sequences from different groups. The alignments are constructed by adding homologous family members to the more distantly related sequences in reference 1. Reference 4 contains alignments of up to 20 sequences including N/C-terminal extensions (up to 400 residues), and Reference 5 consists of alignments including internal insertions (up to 100 residues) [2].

We analyzed the tabu search results from two aspects. The very first set of tests was aimed at to verify the efficiency of our techniques of generating the neighborhood. The techniques are: Swapping (SWP), Node Insertion (NI), Branch Insertion (BI) and Distance Variation (DV). For each neighborhood technique, we ran an extensive set of tests on all the datasets provided by BAliBASE, and computed the scores. The scores using tabu search with each neighborhood generation technique are shown in Table 1. The Number of Test Cases in Reference 1, Reference 2, Reference 3, Reference 4 and Reference 5 are respectively 82, 23, 12, 12 and 12.

One can see in Table 1 that all the neighborhood generation techniques perform well in average for all the reference sets. However, it seems that Branch Insertion and Distance Variation give the best results for all the sequences of Reference 2, Reference 3, Reference 4 and Reference 5. Node insertion gives best results for sequences of Reference 1. We can see that, for all the datasets provided by BAliBASE, Swapping is not the adequate neighborhood technique. This can be explained by the nature of the neighbors generated by a certain technique.

Table 1. Results given by tabu search using four neighborhood generation techniques on the BAliBASE benchmark database

Neighbor-hood	Reference 1	Reference 2	Reference 3	Reference 4	Reference 5	Average
SWP	90.0	93.0	76.3	87.4	85.1	86.36
NI	**90.1**	90.0	78.5	85.6	93.3	87.50
BI	90.05	**93.8**	80.7	**93.7**	**97.9**	**91.23**
DV	90.0	93.5	**82.0**	91.8	95.1	90.48

For the Swapping technique, the neighbors have the same topology, so they are not very different and this will not give more amelioration of the alignment score. For the rest of techniques, the neighbors have not the same topology, but Branch insertion and Distance variation techniques seem to generate more complex guide trees, and this will give more chances to explore different solution spaces and thus, ameliorate the alignment score.

In order to verify the efficiency of our algorithm, we performed another set of tests where the results of our tabu search algorithm using a certain neighborhood technique is compared to other MSA tools. For each references set, we use the adequate neighborhood generation technique which gives the best results, and compare it to the most competitive MSA tools in the literature, such as CLUSTALW 1.83 [28], SAGA [19], MUSCLE [6], ProbCons [5], T-Coffee [20], SPEM [30], PRALINE [27], IMSA ([4] and Tabu Search developed by [24] (called in this paper TS-Riaz) . Except for SAGA and TS-Riaz, which are taken from [24], the results of the other programs are taken from the work of Layeb et al. [16].

The results of our method illustrate clearly the effectiveness of using Tabu Search to perform the multiple sequence alignment. As it can be seen in Table 2, our algorithm performs well in all the references sets. Our method gives good results compared to the other MSA tools. In fact, it gives the second best score for the sequences set Reference 4, the third best score for Reference 3 and Reference 5, and it is in the fourth place for the remaining sets, i.e. Reference 1 and Reference 2. We can see in Table 2 that our Tabu search using Branch Insertion neighborhood technique has a good place for three sequences sets over five, i.e. Reference 2, Reference 4 and Reference 5. Using the Distance Variation neighborhood technique gives the third best score for Reference 3 set, and Node Insertion gives the fourth best score for Reference 1. It can be seen overall, that our tabu search method using Branch Insertion neighborhood technique gives in average the second best score compared to the other algorithms studied in the paper.

Table 2. Results given by Tabu Search using neighborhood techniques compared with other methods on the BAliBASE benchmark database.

Method	Reference 1	Reference 2	Reference 3	Reference 4	Reference 5	Average
CLUSTALW	85.8	93.3	72.3	83.4	85.8	84.12
SAGA	82.5	**95.4**	77.7	78.0	86.8	84.08
MUSCLE	**90.3**	64.4	**82.2**	91.8	**98.1**	85.36
ProbCons	90.0	**94.0**	**82.3**	90.9	**98.1**	91.06
T-Coffee	86.8	**93.9**	76.7	92.1	94.6	88.82
SPEM	**90.8**	93.4	81.4	**97.4**	97.4	**92.08**
PRALINE	**90.4**	94.0	76.4	79.9	81.8	84.5
IMSA	83.4	92.1	78.6	73.0	83.6	82.14
TS-Riaz	76.0	88.9	71.5	77.3	90.5	80.84
TS-SWP	90.0	93.0	76.3	87.4	85.1	86.36
TS-NI	**90.1**	90.0	78.5	85.6	93.3	87.50
TS-BI	90.05	**93.8**	80.7	**93.7**	**97.9**	**91.23**
TS-DV	90.0	93.5	**82.0**	91.8	95.1	90.48

5 Conclusion

In this paper we have demonstrated the efficiency of using tabu search to align multiple sequences. Our algorithm uses several neighborhood generation techniques. To evaluate our approach, we have used BAliBASE benchmark. Firstly, we studied different techniques to produce the neighborhood, then we compared our algorithm to the most recent and competitive MSA tools. We have observed through experiments on BAliBASE that for Reference 1 and Reference 2, the alignments generated by our method are encouraged. For the remaining references, tabu search performs better than most of the other methods studied in this paper.

There are several issues for future work. First, tabu search comes with a number of parameters that can be experimented with to observe the respective effect on the search process. The parameters like tabu list size, tabu tenure, termination criteria, and neighborhood size can have a direct influence on the quality of the final alignment. Further studies are needed to test different scoring schemes and tabu search features.

References

1. Abbas, A., Holmes, S.: Bioinformatics and management science: some common tools and techniques. Operations Research 52(2), 165–190 (2004)
2. Bahr, A., Thompson, J.D., Thierry, J.C., Poch, O.: BAliBASE (benchmark alignment database): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res. 29(1), 323–326 (2001)
3. Blum, C., Valles, M.Y., Blesa, M.J.: An ant colony optimization algorithm for DNA sequencing by hybridization. Computers and Operations Research 38, 3620–3635 (2008)
4. Cutello, V., Nicosia, G., Pavone, M., Prizzi, I.: Protein multiple sequence alignment by hybrid bio-inspired algorithms. Nucleic Acids Research 39(6), 1980–1990 (2010)
5. Do, C., Mahabhashyam, M., Brudno, M., Batzoglou, S.: ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15(2), 330–340 (2005)
6. Edgar, R.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004)
7. Feng, D., Doolittle, R.: Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution 24(4), 351–360 (1987)
8. Glover, F., Laguna, M.: Tabu Search. Kluwer Academic Publishers, Boston (1997)
9. Glover, F., Taillard, E., de Werra, D.: A user's guide to tabu search. Ann. Oper. Res. 41, 3–28 (1993)
10. Gotoh, O.: Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J. Mol. Biol. 264, 823–838 (1996)
11. Gupta, S.K., Kececioglu, J.D., Schaffer, A.A.: Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. J. Comp. Biol. 2(3), 459–472 (1995)
12. Higgins, D.G., Taylor, W.R.: Multiple sequence alignment, Protein Structure Prediction -Methods and Protocols. Humana Press (2000)
13. Kemena, C., Notredame, C.: Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25, 2455–2465 (2009)
14. Kim, J., Pramanik, S., Chung, M.J.: Multiple sequence alignment using simulated annealing. Comp. Applic. Biosci. 10(4), 419–472 (1994)
15. Lalwani, S., Kumar, R., Gupta, N.: A review on particle swarm optimization variants and their applications to multiple sequence alignments. Journal of Applied Mathematics and Bioinformatics 3(2), 87–124 (2013)
16. Layeb, A., Selmane, M., Bencheikh ELhoucine, M.: A new greedy randomized adaptive search procedure for multiple sequence alignment. International Journal of Bioinformatics Research and Applications (2011)
17. Lightner, C.: A Tabu Search Approach to Multiple Sequence Alignment. Ph.D. thesis, North Carolina State University, Raleigh, North Carolina (2008)

18. Lipman, D., Altschul, S., Kececioglu, J.: A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. 86, 4412–4415 (1989)
19. Notredame, C., Higgins, D.G.: SAGA: Sequence alignment by genetic algorithm. Nucl. Acids Res. 24, 1515–1524 (1996)
20. Notredame, C., Higgins, D., Heringa, J.: T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217 (2000)
21. Notredame, C., Holmes, L., Higgins, D.: COFFEE: an objective function for multiple sequence alignments. Bioinformatics 14(5), 407–422 (1998)
22. Orobitg, M., Guitaro, F., Cores, F., Llados, J., Notredame, C.: High performance computing improvements on bioinformatics consistency-based multiple sequence alignment tools (2014), http://dx.doi.org/10.1016/j.parco.2014.09.010
23. Reinert, K., Stoye, J., Will, T.: An iterative method for faster sum-of-pairs multiple sequence alignment. Bioinformatics 16, 808–814 (2000)
24. Riaz, T., Wang, Y., Li, K.: Multiple sequence alignment using tabu search. In: Proceeding of Asia-Pacific Bioinformatics Conference (APBC 2004), pp. 1–10 (2004)
25. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4), 406–425 (1987)
26. Shyu, C., Sheneman, L., Foster, J.: Multiple sequence alignment with evolutionary computation. Genetic Programming and Evolvable Machines 5, 121–144 (2004)
27. Simossis, V., Heringa, J.: PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res. 33, 289–294 (2005)
28. Thompson, J., Higgins, D., Gibson, T.: ClustalW: improving the sensitivity of progressive multiple sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680 (1994)
29. Yokoyama, T., Watanabe, T., Taneda, A., Shimizu, T.: A web server for multiple sequence alignment using genetic algorithm. Genome Informatics, 12, 382–383 (2001)
30. Zhou, H., Zhou, Y.: SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 21, 3615–3621 (2005)

Information Technology: Text and Speech Processing

Noise Robust Features Based on MVA Post-processing

Mohamed Cherif Amara Korba[1,2(✉)], Djemil Messadeg[3], Houcine Bourouba[2], and Rafik Djemili[4]

[1] Mohammed Cherif Messaadia University, Souk-Ahras, Algeria
[2] PI:MIS Laboratory, May 8, 1945 University, Guelma, Algeria
[3] LASA Laboratory, Badji Mokthar University, Annaba, Algeria
[4] August 20, 1955 University, Skikda, Algeria
{amara_korba_cherif,messadeg,bourouba2004
rafik_djemili}@yahoo.fr

Abstract. In this paper we present effective technique to improve the performance of the automatic speech recognition (ASR) system. This technique consisting mean subtraction, variance normalization and application of temporal auto regression moving average (ARMA) filtering. This technique is called MVA. We applied MVA as post-processing stage to Mel frequency cespstral coefficients (MFCC) features and Perceptual Linear Prediction (RASTA-PLP) features, to improve automatic speech recognition (ASR) system.

We evaluate MVA post-processing scheme with aurora 2 database, in presence of various additive noise (subway, babble because, exhibition hall, restaurant, street, airport, train station). Experimental results demonstrate that our method provides substantial improvements in recognition accuracy for speech in the clean training case. We have completed study by comparing MFCC and RSTA-PLP After MVA post processing.

1 Introduction

Most speech recognition systems are sensitive to the nature of the acoustical environments within which they are deployed. The performance of ASR systems decreased dramatically when the input speech is corrupted by various kinds of noise sources. It is quite significant when the test environment is different from the training environment.

In the last two decades, substantial efforts have been made and also number of techniques have been presented to cope with this issue improve the ASR performance. Unfortunately these same algorithms frequently do not provide significant improvements in more difficult environments.

MFCC and RASTA-PLP have served as very successful front-ends for the Hidden Markov Model (HMM) based speech recognition. Many speech recognition systems based on these front-ends have achieved a very high level of accuracy in clean speech environment [14], [15]. However, it is well-known that MFCC is not robust enough in noisy environments, which suggests that the MFCC still has insufficient sound representation capability, especially at low signal-to-noise-ratio (SNR).

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 155–166, 2015.
DOI: 10.1007/978-3-319-19578-0_13

This paper presents noise-robust technique that is simple and effective. The technique post-processing speech features using MVA [10],[11],[12]. The advantage of this technique, it makes no change to the recognition system, it does not change the size of the space, it can be applied on any acoustic feature. it has been shown in [10] and [11] the efficacy of this technique on the database Aurora 2.0 and Aurora 3.0.

This paper is organized as follows: in section 2, we describe MVA post-processing technique, in section3, we show a graphical comparison between different features, in section 4, we present experimental result and in section 5 the work is concluded.

2 Definition and Analyze of MVA Post-Processing Technique

2.1 Definition of MVA Post-Processing Technique

In this part, we describe different steps of development of MVA post-processing technique, Figure 1 provided a block diagram.

For a given utterance, we represent the data by matrix C whose element $C_d(t)$ is the dth component of the feature vector at time t, $t = 1 \dots T$, the number of frames in the utterence and $\mathrm{d} = 1 \dots \mathrm{D}$, the dimension of the feature space, in other words, each column of C represents a time sequence.

$$\begin{bmatrix} C_1(1) & \cdots & C_1(T) \\ \vdots & \ddots & \vdots \\ C_d(1) & \cdots & C_d(T) \end{bmatrix} \tag{1}$$

The first step we application mean subtraction (MS) [6], [7] defined by:

$$\bar{C}_d = C_d(t) - \mu_d \tag{2}$$

Where μ_d is mean vector estimated from data and $\bar{\mathrm{C}}_{\mathrm{d}}$ is the subtracted feature.

$$\mu_d = \frac{1}{T}\sum_{t=1}^{T} C_d(t) \tag{3}$$

MS is an alternate way to high-pass filter cepstral coefficients, it force the average values of cepstral coefficients to be zero in both the training and testing domains. it also removes time-invariant distortions introduced by the transmission channel and recording device.

The second step is Variance normalization (VN) [8], [9] defined by:

$$\tilde{C}_d = \frac{\bar{C}_d(t)}{\sqrt{\sigma_d}} \tag{4}$$

Where σ_d is variance vector estimated from data.

$$\sigma_d = \frac{1}{T}\sum_{t=1}^{T}(C_d(t) - \mu_d)^2 \tag{5}$$

The third step is processing by a mixed auto-regression moving average (ARMA) filtering. In this study we have used two types of ARMA filters: Non Causal ARMA Filter defined by

$$\check{C}_d(t) = \begin{cases} \frac{\sum_{i=1}^{M} \tilde{C}_d(t-i) + \sum_{j=0}^{M} \tilde{C}_d(t+j)}{2M+1} & if\ M < t \le T - M \\ \tilde{C}_d(t) & Otherwise \end{cases} \tag{6}$$

and Causal ARMA Filter defined by :

$$\check{C}_d(t) = \begin{cases} \frac{\sum_{i=1}^{M} \check{C}_d(t-i) + \sum_{j=0}^{M} \tilde{C}_d(t+j)}{2M+1} & if\ M < t \leq T \\ \tilde{C}_d(t) & Otherwise \end{cases} \quad (7)$$

where M is the order of ARMA filter.

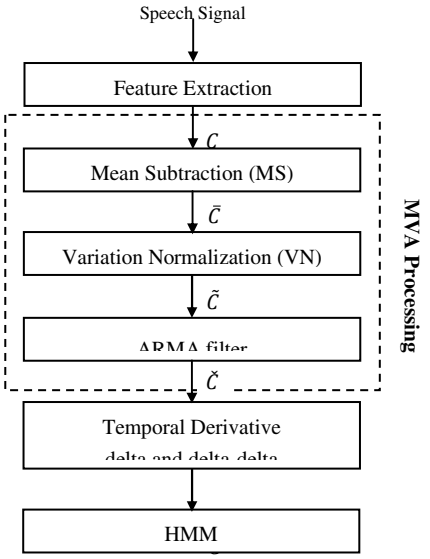


Fig. 1. Block diagram of MVA post-processing technique

In all our experiments, the performances of ASR system are enhanced by adding time derivatives to the basic static parameters for different features. The delta coefficients are computed using the following regression formula:

$$\Delta(t) = \frac{\sum_{b=1}^{B} d\left(\check{C}_d(b+1) - \check{C}_d(b-1)\right)}{2\sum_{b=1}^{B} b^2} \quad (8)$$

Where $\Delta(t)$ is the delta coefficient computed in terms of the corresponding static coefficients $\tilde{C}_d(t-B)$ to $\tilde{C}_d(t+B)$. The same formula is applied to the delta to obtain acceleration coefficients.

2.2 Effect of Normalization and ARMA Filter on Acoustic Features

In Fig. 2 and fig. 3 the time sequences of C0 and C1 are plotted for both features RASTA-PLP and MFCC of the utterance of digit string "98Z7437" corrupted by different levels of additive subway noise from the Aurora 2.0 database. For both RASTA-PLP and MFCC features, we see enormous differences between the plots of the clean case and the more noisy case. In particular, the clean and noisy plots have quite a different average value and dynamic range.

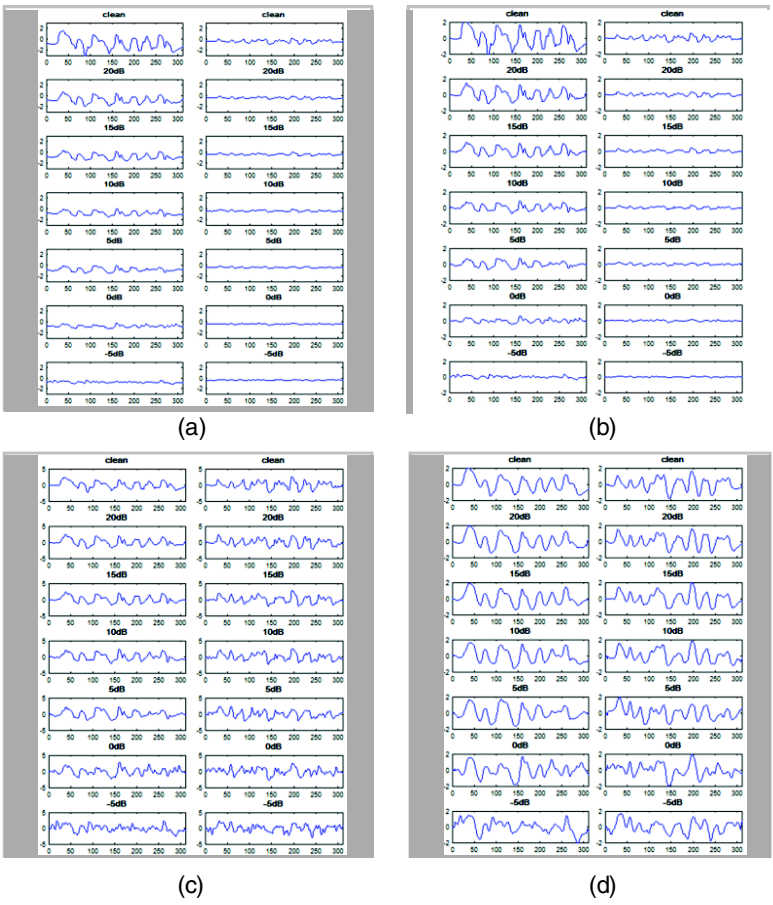


Fig. 2. The time sequence of C0 and C1 coefficients of RASTA-PLP features for the digit string "98Z7437" corrupted by additive subway noise, (a) RASTA-PLP features, (b) time sequence of RASTA-PLP + MS, (c) time sequence of RASTA-PLP + MS + VN, (d) time sequence of RASTA-PLP + MVA.

After MS and VN is applied, the difference between the clean and noisy cases are made much less severe After MS and VN is applied, the differences between the clean and noisy cases are made much less severe. Still, however, some differences remain between the clean and noisy cases. We notice in particular the case of C1, that after the application of MS and VN, the time sequences in noisy speech show spurious spikes relative to the clean case. In order to further reduce differences, we apply ARMA filtering which smoothes out the sequences thus making them more similar to each other. We remark that, the effects of noise on the MVA features are less severe for both MFCC and RASTA-PLP features.

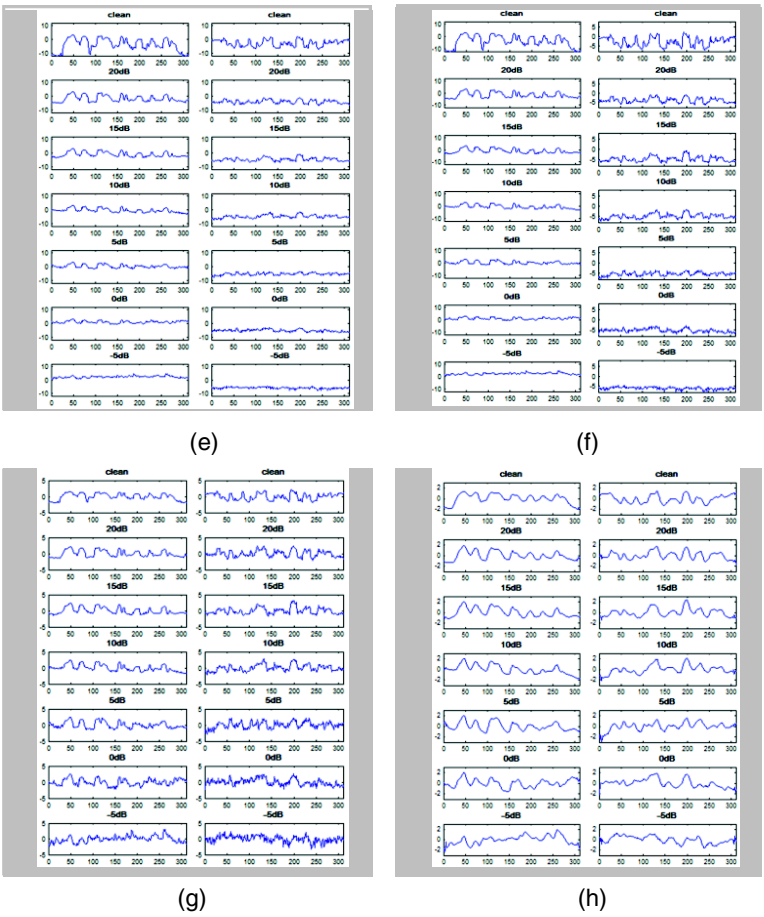
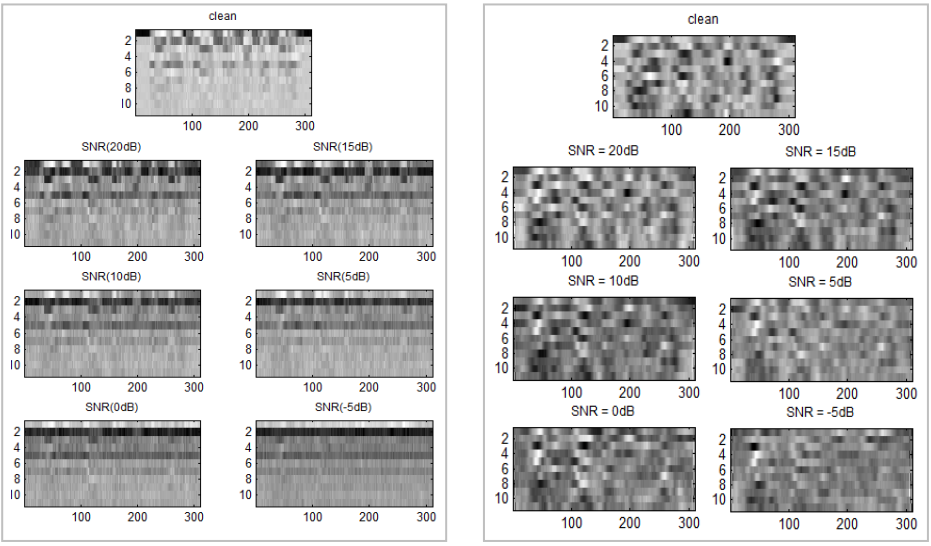


Fig. 3. The time sequence of C0 and C1 coefficients of MFCC features for the digit string "98Z7437" corrupted by additive subway noise, (e) MFCC features, (f) time sequence of MFCC + MS, (g) time sequence of MFCC + MS + VN, (h) time sequence of MFCC + MVA

3 Graphical Comparison between the Different Features

Fig. 4 shows a sample comparison between baseline MFCC features and corresponding MFCC MVA post-processing features for the digit string "98Z7437" corrupted with Subway noise at different levels of noise (clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB). As standard in MFCC, a window size of 25 ms with an overlap of 10 ms was chosen, and Cepstral features were obtained from DCT of log-energy over 23 Mel-scale filter banks.

The degradation of spectral features for baseline MFCC features in the presence of noise is evident; whereas MFCC with MVA post-processing features obtained with No Causal ARMA filter prevail at elevated noise levels. For SNR $\leq$ 0dB we can see clearly that MFCC with MVA is better noise robustness than MFCC baseline features.



(a) (b)

Fig. 4. (a) Baseline MFCC features for the digit string "98Z7437" corrupted by subway noise, (b) MFCC with MVA post-processing features for the digit string "98Z7437" corrupted by subway noise. (No causal ARMA filter used, filter order = 5)

3.1 Speech Features Description

This part contains a short description of the most widely used acoustic features in automatic speech recognition. Many of current ASRs are based on Mel frequency cepstral coefficients MFCC [5] or RASTA-PLP coefficients [3],[4]. They operate efficiently in the clean environment, by against the performances of ASR decreases dramatically in presence of noise. To remedy this problem, we introduced a post-processing stage to improve their performances without bringing changes in their structures. Table 1 shows the configuration of MFCC and RASTA-PLP features used for experiences.

4 Experiments

We first describe in detail the Aurora 2 database, then, we present experimental results that are intended to show the contribution of MVA post-processing technique for both acoustic features MFCC and RASTA-PLP in the presence of large variety of additive noise. We determine type and order of ARMA filter that gives the best speech accuracy for acoustic features used (MFCC and RASTA-PLP).

4.1 Description of Aurora 2 Database

Our speech recognition experiments were conducted using the Aurora 2 database and task [2]. The Aurora task [2] has been defined by the European Telecommunications Standards (ETSI) to standardize a robust feature extraction technique for a distributed speech recognition framework.

The Aurora 2 database is a subset of the TIDigits, which contains a set of connected digit utterances spoken in English; while the task consists of the recognition of the connected digit utterances interfered with real noise artificially added in a wide range of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB, 20dB and Clean) and the channel distortion is additionally included in Set C. Noise signals are recorded at different places including suburban train, babble, car, exhibition hall, restaurant, street, airport and train station.

Two training modes are defined, training on clean data only and training on clean as well as noisy data (multi-condition). For the first mode, training data contain 8440 clean utterances produced by 55 male and 55 female adults. For the multi-condition training, 8440 utterances from TIDigits training parts are equally split into 20 subsets with 422 utterances in each subset. Four types of noise, Suburban train, babble, car, and exhibition hall noises are added to 20 subsets at 5 different SNRs (5dB, 10dB, 15dB, 20 dB and Clean).

The testing data consist of 4004 utterances from 52 male and 52 female speakers in the TIDigits test part are divided into four subsets with 1001 utterances in each. One noise is added to each subset at SNRs of 20 to -5 dB in decreasing steps of 5 dB after speech and noise are being filtered with the G. 712. Three test sets are defined as below:

Test Set A: four types of noise, babble, car, suburban train, and exhibition hall are added to the four subsets of utterances to produce 28028 utterances (4x7x1001 utterances). This set leads to a high match of training and test data as it contains the same noises as used for the multi-condition training mode.

Test Set B: the other type of noise, street, restaurant, airport and train station, are added to the four subsets of utterances to produce 28028 utterances (4x7x1001 utterances), similar to test A.

Test Set C: two types of noise, suburban train and street, are individually added to two of the four subsets of utterances to produce 14014 utterances (2x7x1001 utterances). Speech and noise are filtered with the MIRS frequency characteristic before adding.

In this study we used two sets of tests, Test Set A and Test Set B. for all experiments HMM baseline system is trained in clean condition.

4.2 The HTK Recognizer

For the baseline system, the training and recognition tests used the HTK recognition toolkit [1], which followed the setup originally defined for the ETSI Aurora evaluations.

Each digit was modeled as a left to right continuous density HMM with 16 states with each state having 3 mixtures. Two pause models, silence "sil" and short pause "sp", were defined. The "sil" model had three states with six Gaussian mixtures per state. The "sp" model had one state with six Gaussian mixtures.

Script files provided with the Aurora 2 database for the purpose of training and testing a HTK based recognizer were used in the evaluation of the front-ends. The version of HTK used was HTK 3.3. We used the RASTA-PLP implementation that is valuable at [13], we used the version of conventional MFCC processing implemented as part of HTK platform. Configurations of RASTA-PLP and MFCC features used in our experiments are given by the table 1.

Table 1. features parameters used for experimental analysis

Configuration features	MFCC	RASTA-PLP
Frame length (ms)	25	25
Frame shift (ms)	10	10
Pre-emphasis coefficient	0.97	NO
Analyses window	Hamming	NO
frequency range	64 – 4000 Hz	0 – 4000 Hz
No. Mel filterbanks	23	/
LPC Model order	/	11
Rasta filter	/	do
Appended log frame energy	yes	yes
Appended features	Δ + Δ Δ	
Δ window (frames)	±4	±4
ΔΔ window (frames)	±1	±1
Feature dimension	39	39

4.3 Analyses

The tables below were done to determine type and order of ARMA filter that gives the best recognition accuracy for each acoustic feature. Tables show the contribution importance of order of filter on the performances of ASR system.

For all our experiments, best results have been obtained with the non-causal ARMA filter for both acoustic features. We varied the order of the filter until 9, the best performances of the system have been obtained with order M = 6.

Table 2. Comparison of different type and order ARMA filters, word accuracy Rasta-PLP, Test speech average over SNR (clean, 20, 15, 10, 5, 0, -5dB)

	Filter Order				
Filter Type	2	3	4	5	6
Causal ARMA filter	67.52	69.09	69.06	68.94	**68.98**
Non Causal ARMA filter	68.33	68.63	69.74	69.66	**70.46**

Table 3. Comparison of different type and order ARMA filters, word accuracy MFCC, Test speech average over SNR (clean, 20, 15, 10, 5, 0 -5dB)

	Filter Order				
Filter Type	2	3	4	5	6
Causal ARMA filter	67.57	66.84	67.74	69.35	**70.55**
Non Causal ARMA filter	67.67	69.10	69.53	70.05	**71.40**

4.4 Performance of MVA pPost-processing

In this section we describe the recognition accuracy obtained using MVA post-processing for MFCC and RASTA-PLP features, under various noise conditions at different SNR levels (Clean, 20, 15, 10, 5, 0, -5dB).

In figure 5, remarkably improvements have been achieved up to 20% compared to RASTA-PLP features without any normalization, up to 10% to features with MSVN normalization and up to 5% to features with MS normalization.

In figure 6, Substantial improvements have been achieved up to 25% compared to MFCC features without any normalization, up to 15% to features with MS + VN normalization and up to 8% to features with MS normalization.

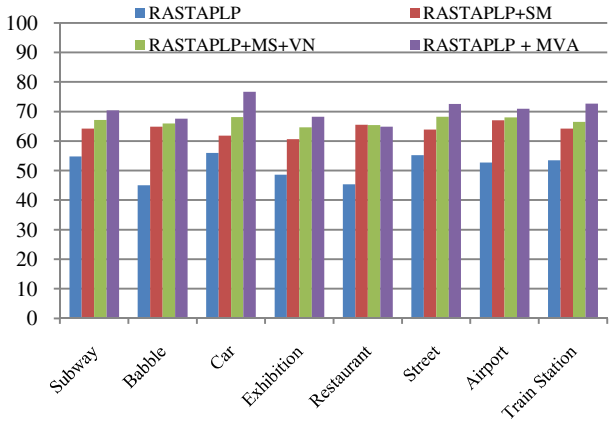


Fig. 5. Comparison of recognition accuracy for different RASTA-PLP features configuration (MVA: use non causal ARMA filter, M = 6), the recognition accuracy is calculated on an average of 7 SNR levels. (Clean, 20dB, 15dB, 10dB, 5dB, 10dB, 5dB, 0dB, -5dB).

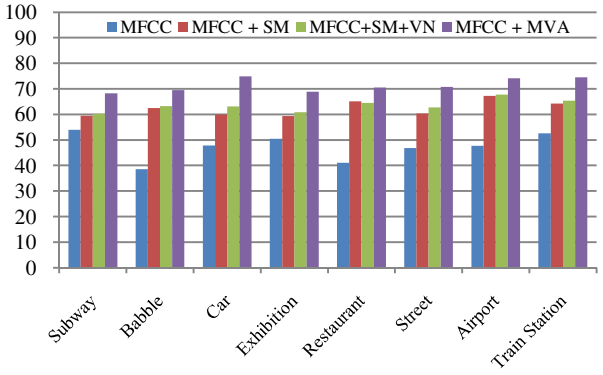


Fig. 6. Comparison of recognition accuracy for different MFCC features configuration (MVA: use non causal ARMA filter, M = 6), the recognition accuracy is calculated on an average of 7 SNR levels. (Clean, 20dB, 15dB, 10dB, 5dB, 10dB, 5dB, 0dB, -5dB).

Figure 7 shows a comparison between MFCC features and Rasta-PLP features, in the presence of stationary noise subway, street and car the RASTA-PLP features are more efficient compared to MFCC features, but in the presence of noises majority babble, suburban train, exhibition hall, restaurant, airport and the train station the MFCC coefficients provide best performance to ASR system.

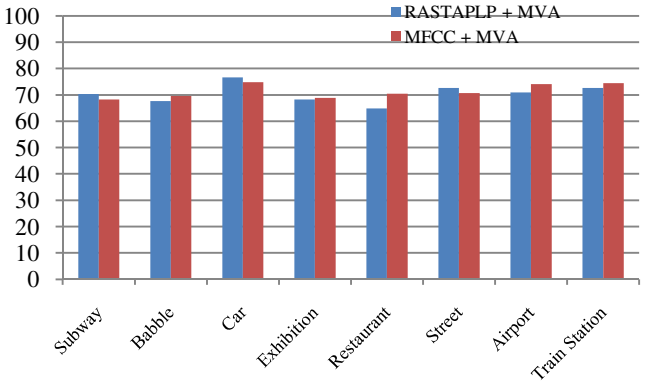


Fig. 7. Comparison of recognition accuracy for MFCC + MVA features with RASTA-PLP + MVA features (MVA: use non causal ARMA filter, M = 6 for both types of features), the recognition accuracy is calculated on an average of 7 SNR levels. (Clean, 20dB, 15dB, 10dB, 5dB, 10dB, 5dB, 0dB, -5dB).

5 Conclusions

In this paper, we introduce MVA technique to MFCC and RASTA-PLP features to improve the noise robustness of speech features. We have shown that normalization techniques followed by ARMA filter are vital for conditions with major mismatch between training and test condition.

The experimental results show that application of MVA to the Aurora 2 database can provide further robustness to noise for various types of features, and higher accuracy rates can be thereby achieved.

Acknowledgment. This work was supported by PI:MIS laboratory of Guelma University. The authors would like to thank Professor A. Boukrouche and H. Doghmane for Helpful discussion.

References

1. Young, S., et al.: The HTK Book Version 3.3 (2005)
2. Hirsch, H.G., Pearce, D.: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. In: Proc. ISCA ITRW ASR (2000)
3. Hermansky, H.: Perceptual linear prediction analysis of speech. J. Acoust. Soc. Am. 87(4), 1738–1752 (1990)
4. Hermansky, H., Morgan, N.: RASTA processing of speech. IEEE. Trans. Speech Audio Process. 2(4), 578–589 (1994)
5. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, and Signal Processing 28(4), 357–366 (1980)
6. Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America 55, 1304–1312 (1974)
7. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust., Speech, Signal Process. 29(2), 254–272 (1981)
8. Jain, P., Hermansky, H.: Improved mean and variance normalization for robust speech recognition. In: IEEE Int. Conf. Acoust., Speech and Signal Processing (May 2001)
9. Cook, G.D., Kershaw, D.J., Christie, J.D.M., Seymour, C.W., Waterhouse, S.R.: Transcription of broadcast television and radio news: the 1996 abbot system. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Munich, Germany (1997)
10. Chen, C.-P., Bilmes, J., Kirchhoff, K.: Low-resource noise-robust feature post-processing on Aurora 2.0. In: Proc. Int. Conf. Spoken Lang. Process. (ICSLP), pp. 2445–2448 (2002)
11. Chen, C.-P., Filali, K., Bilmes, J.: Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases. In: Proc. Int. Conf. Spoken Lang. Process. (ICSLP), pp. 241–244 (2002)
12. Chen, C.-P., Bilmes, J.: MVA processing of speech features Dept. Elect. Eng., Univ. Washington, Seattle, WA, Tech. Rep. UWEETR- 2003-0024 (2003), `http://www.ee.washington.edu/techsite/papers`

13. Ellis, D.: PLP and RASTA (and MFCC, and inversion) in MATLAB using melfcc.m and invmelfcc.m (2006), http://labrosa.ee.columbia.edu/matlab/rastamat/
14. Stuttle, M.N., Gales, M.J.F.: A Mixture of Gaussians Front End for Speech Recognition. In: Eurospeech 2001, Scandinavia, pp. 675–678 (2001)
15. Potamifis, J., Fakotakis, N., Kokkinakis, G.: Improving the robustness of noisy MFCC features using minimal recurrent neural networks. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, vol. 5, pp. 271–276 (2000)

Arabic Texts Categorization: Features Selection Based on the Extraction of Words' Roots

Said Gadri[1(✉)] and Abdelouahab Moussaoui[2]

[1] Department of ICST, University of M'sila, 28000, Algeria
kadri.said28@yahoo.fr
[2] Department of Computer Sciences, University Farhat Abbes of Setif, Setif, 19000, Algeria
moussaoui.abdel@gmail.com

Abstract. One of methods used to reduce the size of terms vocabulary in Arabic text categorization is to replace the different variants (forms) of words by their common root. The search of root in Arabic or Arabic word root extraction is more difficult than other languages since Arabic language has a very different and difficult structure, that is because it is a very rich language with complex morphology. Many algorithms are proposed in this field. Some of them are based on morphological rules and grammatical patterns, thus they are quite difficult and require deep linguistic knowledge. Others are statistical, so they are less difficult and based only on some calculations. In this paper we propose a new statistical algorithm which permits to extract roots of Arabic words using the technique of n-grams of characters without using any morphological rule or grammatical patterns.

Keywords: Root extraction · Information retrieval · Bigrams technique · Arabic morphological rules · Feature selection

1 Introduction

Arabic is one of the oldest and the most used language in the world, it is spoken by over 300 million people in Arabic world, and used by more than 1.7 billion Muslims over the world because it is the language of the Holy Quran, here we can distinguish two types of Arabic; a more classical language, as found in the Holy Quran or poetry, a standardized modern language, and regional dialects [1]. We note also that Arabic language is a semitic language [2, 3] based on 28 cursives letters written from right to left.

The word in Arabic is formed of the root part and some affixes (antefixes, prefixes, infixes, suffixes) that form the word (سألتمونيها Saaltmwnyha). The Arabic root extraction is a very difficult task which is not the case for other languages as English or French, because Arabic is a very rich language with a very difficult structure and complex morphology. Arabian linguists show that all nouns and verbs of Arabic language are derived from a set of roots containing about 11347 roots; more than 75 % of them are trilateral roots [4].

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 167–180, 2015.
DOI: 10.1007/978-3-319-19578-0_14

There are many applications based on the roots of words in Arabic processing such as: text's classification, text summarizing, information retrieval, data and text mining. [5,6].

The Arabic words ' roots can be classified according to the vowels letters (أ ، و ، ي a, w, y) into two types [7], strong roots that do not contain any vowel (فتح ، خرج ، ذهب go, come out, open), vocalic roots that contain at least one vowel (وعد ، أوى shelter, promise). Arabic roots can be further classified according to the number of their characters into four types: Trilateral roots which form most words in Arabic language [4] (e.g., خرج ، كتب ، علم know, write, come out), Quadrilateral roots (e.g., دحرج ، طمأن roll, assure), Quinquelateral roots (e.g., انطلق ، اقتصد ، انكسر broken, economize, start) and Hexalateral roots (استعمل ، استحسن اقشعّر use, enjoy, tremble).

There are two classes of methods used to extract the roots of Arabic words, the first class is based on morphological rules, so its methods simulate the same process of an expert linguist during his analysis of a given Arabic word [1], [8,9,10,11], which make the process of extracting root difficult and complex because of the diversity of morphological formulas and the multiplicity of words forms for the same root when changing the original characters position in the word (e.g., علم ، عالم ، علوم ، عوالم ، معالم know, scientist, sciences, worlds, landmarks) [12,13]. The second class is formed of statistical methods which are simple, fast, and do not require any morphological rules but some calculations [14,15, 16,17, 18,19,20].

In this paper, we propose a new statistical method which permits to extract roots of Arabic words using the approach of n-grams of characters without using any morphological rule. The paper is organized as follows: the first section is a general introduction to the field of study. The second section presents some related works, so we review some papers that treat the problem of extraction of Arabic word's roots. In the third section we introduce our new algorithm. The fourth section presents the experiments that we have done to test our new method and also presents the obtained results. In the last section we conclude our work by summarizing our realized work and giving some ideas to improve it in the future.

2 Related Works

Many researchers proposed some algorithms to extract Arabic words roots, some of these algorithms are based on morphological rules. Thus, they are called morphological methods. Others do not use any morphological rule but some statistical calculations, so they are called statistical algorithms.

In the first class of algorithms, we can note the following: [9], [21] Khoja's roots extractor removes the longest suffix and prefix. It then matches the remaining word with verbal and noun patterns, to extract the root. The roots extractor makes use of several linguistic data files such as a list of all diacritics, punctuation characters, definite articles, and stop words [22,23,24,25]. [13] Propose a linguistic approach for root extraction as a preprocessing step for Arabic text mining. The proposed approach is composed of a rule-based light stemmer and a pattern-based infix remover. They propose an algorithm to handle weak, eliminated-long vowel, hamzated and geminated words. The accuracy of the extracted roots is determined by comparing

them with a predefined list of 5,405 trilateral and quadrilateral roots. The linguistic approach performance was tested on texts' collection consists of eight categories, the author achieved a success ratio about 73.74%. [26] Presents a new Arabic root extractor that tries to assign a unique root for each Arabic word without having an Arabic roots list, a word patterns list, or the list of Arabic prefixes and suffixes. The algorithm predict the letters positions that may form the word root one by one, using rules based on the relations between the Arabic word letters and their placement in the word. This algorithm consists of two parts, the first part gives the rules that distinguish between the Arabic definite letter "الـ AL, La" and the original word letters "الـ". The second part segments each word into three parts and classifies its letters according to their positions. The author tested her proposed algorithm using the Holy Quran words and obtained an accuracy of 93.7% in root extracting process.

In the second class of algorithms we can note the following: [14] Developed a root extraction algorithm which does not use any dictionary, their algorithm categorizes all Arabic letters according to six integer weights, ranging from 0 to 5, as well as the rank of the letter which is determined by the position this letter holds in a word. The weight and rank are multiplied together, and the three letters with the smallest product constitute the root of the word. We note that [14] did not explain on what basis did it use such ranking or weighting. [10] Proposes an algorithm to extract tri-literal Arabic roots, this algorithm consists of two steps; in the first step they eliminate stop words as well prefixes and suffixes. In the next step, they remove the repeated word's letters until only three letters are remained, and then they arrange these remaining letters according to their order in the original word, which form the root of the original word. The obtained results were very promising and give an accuracy of root's extraction over than 73%. [27] Propose a new way to extract the roots of Arabic words using n-grams technique. They used two similarity measures; the "Manhattan distance measurement" and the "Dice's measurement". They tested their algorithm on the Holy Quran and on a corpus of 242 abstracts from the Proceedings of the Saudi Arabian National Computer Conferences. They concluded from their study that combining the n-grams with the Dice's measurement gives better results than using the Manhattan distance measurement. [28] propose a new algorithm to find a system that assigns, for every non vowel word a unique root. The proposed system consists of two modules; the first one consists of analyzing the context by segmenting the words of the sentence into its elementary morphological units in order to extract its possible roots. So, each word is segmented into three parts (prefix, stem and suffix). In the second module, they based on the context to extract the correct root among all possible roots of the word. They validate their algorithm using NEMLAR Arabic writing corpus that consists of 500,000 words, and their proposed algorithm gives the correct root in more than 98% of the training set and 94% of the testing set. [29] Propose a new algorithm which use the n-grams technique. In this technique, both the word and its assumed root are divided into pairs called bi-grams, then the similarity between the word and the root is calculated using equation (1) [30]. This process is repeated for each root in the roots list:

$$S = 2 \times C/(A + B) \quad (1)$$

Where:
A = Number of unique bi-grams in the word (A)
B = Number of unique bi-grams in the root (B)
C = Number of similar unique pairs between the word (A) and the root (B)

To use equation (1) for extracting the word's root, we must have: the word (A) and the potential roots (B) to compare with, then the similarity measuring is conducted by computing the value of (S) between the word (A) and each potential roots (B).

3 The Proposed Algorithm

In our new algorithm, we use also the n-grams technique to extract Arabic words roots, for this purpose, we proceed according to the following steps:

Step 1: we segment the word for which we want to find the root, and all the roots of the list into bigrams (2-grams).

For example if we have the word "**يذهبون**" and a list of six (06) roots (**فتح ، خرج ، ذهب ، وجد ، وهب ، نهب**), we proceed the segmentation step as follows:

W = "يذهبون" ➔ (يذ، يه ، يب ، يو ، ين ، ذه ، ذب ، ذو ، ذن ، هب ، هو ، هن ، بو ، بن ، ون)
R_1 = "فتح" ➔ (فت ،فح ، تح)
R_2 = "خرج" ➔ (خر ، خج ، رج)
R_3 = "ذهب" ➔ (ذه ، ذب ، هب)
R_4 = "وجد" ➔ (وج ، ود ، جد)
R_5 = "وهب" ➔ (وه ، وب ، هب)
R_6 = "نهب" ➔ (نه ، نب ، هب)

Step 2: we calculate the following parameters:

N_W : The number of unique bigrams in the word *w*

N_{R_i} : The number of unique bigrams in the root *Ri*

N_{WR_i} : The number of common unique bigrams between the word *W* and the root *Ri*

$N_{W\bar{R}_i}$: The number of bigrams belonging to the word w and do not belong to the root *Ri*

$$(N_{W\bar{R}_i}=N_W - N_{WR_i})$$

$N_{R_i\bar{W}}$: The number of bigrams belonging to the root *Ri* and do not belong to the word *w*

$$(N_{R_i\bar{W}}= N_{R_i} - N_{WR_i})$$

For the previous example we have:

N_W=18, N_{R_1}=3, N_{R_2}=3, N_{R_3}=3, N_{R_4}=3, N_{R_5}=3, N_{R_6}=3, N_{WR_1}=0, N_{WR_2}=0, N_{WR_3}=3, N_{WR_4}=0, N_{WR_5}=1, N_{WR_6}=1, $N_{W\bar{R}_1} = 18$, $N_{W\bar{R}_2} = 18$, $N_{W\bar{R}_3} = 15$, $N_{W\bar{R}_4} = 18$, $N_{W\bar{R}_5} = 17$, $N_{W\bar{R}_6} = 17$, $N_{R_1\bar{W}} = 3$, $N_{R_2\bar{W}} = 3$, $N_{R_3\bar{W}} = 0$, $N_{R_4\bar{W}} = 3$, $N_{R_5\bar{W}} = 2$, $N_{R_6\bar{W}} = 2$.

Step3: we take only the roots having at least one common bigram with the word *w* ($N_{WR_i} \geq 1$) as candidate roots among the list of all roots in order to reduce the calculation time.

In our previous example, we can take only the roots: R_3 = "ذهب", R_5 = "وهب", R_6 = "نهب" with N_{WR_i}= 3, 1, 1 respectively.

Step4: we calculate the distance $D(w, R_i)$ between the word *W* and each candidate root R_i (R_3, R_5, R_6) according to the following equation :

$$D(w, R_i) = 2 * N_{wR_i} + k * N_{w\bar{R}_i} + k * N_{R_i\bar{w}} \quad (2)$$

Where: *k* is a constant which must take a high value (we put here *k*=100)

For the previous example we obtain:

$D(w, R_3) = 2*3+15*100+0*100 = 1506$

$D(w, R_5) = 2*1+17*100+2*100 = 1902$

$D(w, R_6) = 2*1+17*100+2*100 = 1902$

Step5: in the last step, we assign the root that has the lowest value of distance $D(w, R_i)$ among the candidate roots to the word *W*. it is the required root.

In our example, the root of the word "يذهبون" is "ذهب"

Finally, we note that our new algorithm has the following advantages:

1. Does not require the removal of affixes whose distinction from the native letters of the word is quite difficult.
2. Works for any word whatever the length of the root.
3. Valid for strong roots and vocalic roots which generally pose problems in Arabic during their derivation, because of the complete change of their forms.
4. Does not use any morphological rule nor patterns but simple calculations of distances.
5. Very practical algorithm and easy to implement on machine.

4 Experimentations and Obtained Results

To validate our proposed algorithm, we used three corpus which can be classified according their sizes into: small corpus, middle corpus, and large corpus.
Each one is constituted of many files as indicated below:

1. The file of derived forms (gross words) which contains morphological forms of words derived from many Arabic roots.
2. The file of roots which contains many Arabic roots, we note that these roots are trilateral, quadrilateral, quinquelateral, and hexalateral. We note also that many of them are vocalic roots which contain at least one vowel.

3. The file of golden roots which contain the correct roots of all words present in our corpus (the file in (1)), this golden list was prepared by an expert linguist and used as reference list, i.e., by comparison between the list of obtained roots (extracted by the system) and the reference list (established by the expert), we can calculate the roots extraction accuracy (success ratio).

Table 1. Corpus used in experiments

Corpus	Size of derived words' file	Size of the roots' file	Size of the golden roots' file
Small corpus	50	25	50
Middle corpus	270	135	270
Large corpus	1500	450	1500

Table 2. An example of morphological forms (gross words)

Word	Word	Word	Word	Word
مأخذ	أوامر	باحث	اجتماعات	مأخذ
مؤاخذة	مؤتمر	بحوث	اجتماعيات	مؤاخذة
مؤاخذون	مؤامرة	أبحاث	جموع	مؤاخذون
مؤاخذات	متأمرون	باحثون	جوامع	مؤاخذات
مؤازرة	يأتمرون	باحثات	يجمعون	مؤازرة
مأكل	يأتمرن	ابتهال	يجمعن	مأكل
أكلات	أمرهم	مبتهل	اجتهاد	أكلات

Table 3. An example of trilateral, quadrilateral, quinquelateral, hexalateral roots

Trilateral roots	Quadrilateral roots	Quinquelateral roots	Hexalateral roots
زرع	أكرم	انطلق	استعمل
صنع	أعان	انكسر	استحسن
تجر	أعطى	احتوى	استعان
جمع	حطّم	اقتصد	اخشوشن
نفر	ربّى	اخضرّ	ادهامّ
طار	حاسب	تحدّى	احرنجم
سعل	طمأن	تنازل	اقشعرّ
صدع	زلزل	تدحرج	اطمأنّ

Table 4. Examples of obtained results when segmenting words into bi-grams

Word	N-grams Ng.Frequencies	Nb.Ng (N_W)
يتعلمون	يت يع يل يم يو ين تع تل تم تو تن عل عم عو عن لم لو لن مو من ون 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	28
عالم	عا عل عم ال ام لم 1 1 1 1 1 1	6
كاتب	كا كت كب ات اب تب 1 1 1 1 1 1	6
كتاتيب	كت كا كت كي كب تا تت تي تب ات اي اب تي تب يب 2 1 1 1 1 1 2 2 1 1 1 1	12
اقتصاد	اق ات اص اا اد قت قص قا قد تص تا تد صا صد اد 1 1 1 12 1 1 1 1 11 1 1 1	14
يقصدون	يق يص يد يو ين قص قد قو قن صد صو صن دو دن ون 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	15
استخدم	اس ات اخ اد ام ست سخ سد سم تخ تد تم خد خم دم 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	15
سنستدرجهم	سن سس ست سد سر سج سه سم نس نت ند نر نج نه نم ست سد سر سج سه سم تد تر تج ته تم در دج ده دم رج ره رم جه جم هم 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 12 2 2 2 2 2 2 1 1 1 11 1	30
متذبذب	مت مذ مب مذ مب تذ تب تذ تب ذذ ذب بذ بب ذب 1 12 2 2 2 3 1 1	9
متلألئ	مت مل مأ مل مئ تل تأ تل تئ لأ لل لئ أل أئ لئ 1 12 1 1 2 1 1 1 1 2 1	12
يهزمونهم	يه يز يم يو ين يه يم هز هم هو هن هه هم زم زو زن زه زم مو من مه مم ون وه وم نه نم هم 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 3 1 1 1 2 1 2	23
المتربّي	ال ام ات ار اب آ اي لم لت لر لب لّ لي مت مر مب مّ مي تر تب تّ تي رب رّ ري بّ بي ّي 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	28
المربّون	ال ام ار اب آ او ان لم لر لب لّ لو لن مر مب مّ مو من رب رّ رو رن بّ بو بن ّو ّن ون 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	28
طائرات	طا طئ طر طا طت ائ ار اا ات ئر ئا ئت را رت ات 1 1 1 1 1 2 1 1 1 1 1 1 2	13

Table 5. Examples of obtained results when segmenting roots into bi-grams

Root	N-grams Ng.Frequencies	Nb.Ng (N_{R_i})
كلم	كل كّ كم لّ لم ّم 1 1 1 1 1 1	6
عالج	عا عل عج ال اج لج 1 1 1 1 1 1	6
قصد	قص قد صد 1 1 1	3
اقتصد	اق ات اص اد قت قص قد تص تد صد 1 1 1 1 1 1 1 1 1 1	10
كتب	كت كب تب 1 1 1	3
علم	عل عم لم 1 1 1	3
عمل	عم عل مل 1 1 1	3
خدم	خد خم دم 1 1 1	3
كمل	كم كل مل 1 1 1	3
كمن	كم كن من 1 1 1	3
خمد	خم خد مد 1 1 1	3
درج	در دج رج 1 1 1	3
ذبذب	ذب ذذ ذب بذ بب ذب 1 1 1 3	4
لألأ	لأ لل لأ أل أأ لأ 1 1 1 3	4
هزم	هز هم زم 1 1 1	3
طار	طا طر ار 1 1 1	3
ربّى	رب رّ ري بّ بيّ ي 1 1 1 1 1 1	6
عقد	عق عد قد 1 1 1	3
تأتأ	تأ تت تأ أت أأ تأ 1 1 1 3	4

Table 6. Extraction of some Arabic words roots using our new algorithm

Word	Nearest roots	Nb.Common bi-grams	Distance values	Extracted root	Correct root
يتعلمون	كلم ، عالج ، علم ، عمل ، كمن	3 ، 1 ، 3 ، 2 ، 1	**2506** ، 3202 ، 2806 ، 2902 ، 2704	علم	علم
عالم	كلم ، عالج ، علم ، عمل	1 ، 3 ، 3 ، 2	**306** ، 606 ، 1002 ، 504	علم	علم
كاتب	اقتصد ، كتب	1 ، 3	**306** ، 1402	كتب	كتب
كتاتيب	اقتصد ، كتب ، تأتأ	1 ، 3 ، 1	**906** ، 2002 ، 1402	كتب	كتب
اقتصاد	قصد ، اقتصد ، عقد	3 ، 10 ، 1	**420** ، 1106 ، 1502	اقتصد	اقتصد
يقصدون	قصد ، اقتصد ، عقد	3 ، 3 ، 1	**1206** ، 1906 ، 1602	قصد	قصد
استخدم	اقتصد ، خمد ، خدم	3 ، 2 ، 3	**1206** ، 1404 ، 1906	خدم	خدم
سنستدرجهم	اقتصد ، خدم ، درج ، هزم	1 ، 1 ، 3 ، 1	**2706** ، 3102 ، 3802 ، 3102	درج	درج
متذبذب	كتب ، ذبذب	1 ، 4	**508** ، 1002	ذبذب	ذبذب
متلألئ	عمل ، كمل ، تأتأ ، لألأ	1 ، 1 ، 1 ، 3	1302 ، 1302 ، 1402 ، 1006	لألأ	لألأ
يهزمونهم	كمن ، هزم	1 ، 3	**2402** ، 2006	هزم	هزم
المتربّي	كلم ، عالج ، اقتصد ، كتب ، علم ، ربّى ، طار	1 ، 1 ، 1 ، 1 ، 1 ، 6 ، 1	2902 ، 3202 ، 3602 ، **2212** ، 2902 ، 2902	ربّى	ربّى
المربّون	كلم ، عالج ، علم ، كمن ، ربّى ، طار	1 ، 1 ، 1 ، 1 ، 3 ، 1	2902 ، 3202 ، 2902 ، **2806** ، 2902	ربّى	ربّى
طائرات	اقتصد ، طار	3 ، 1	**1006** ، 2102	طار	طار

Table 7. Obtained results when extracting the words roots

Corpus	Nb.Roots	Nb.Words	Cor. Results	Wr.Results	Suc.Rate	Err.Rate
Small	25	50	49	1	98,00	2,00
Middle	135	270	253	17	94,07	5,93
Large	450	1500	1358	142	90,53	9,47

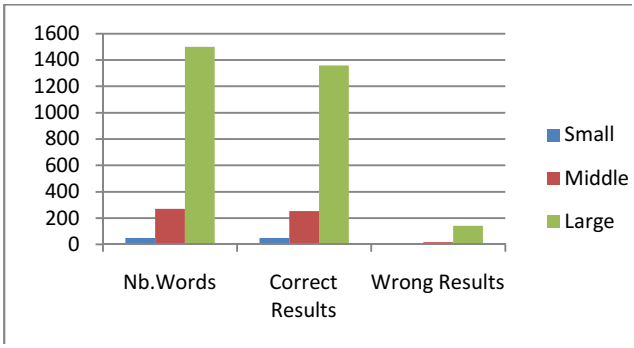


Fig. 1. Correct and wrong results in number of words

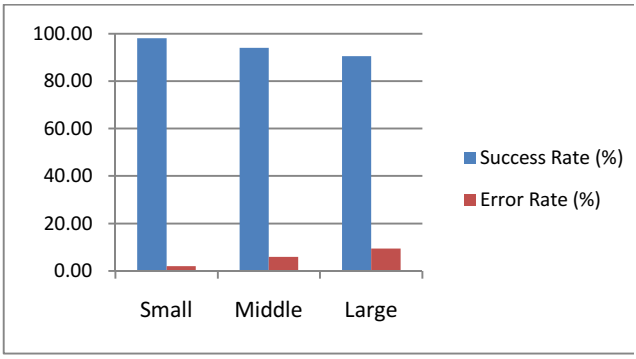


Fig. 2. Calculation of success rate and error rate

5 Comparison with Other Algorithms

To show the effectiveness of our proposed algorithm, we concluded our work by establishing a comparison against other known algorithms. For this purpose, we took a sample words list and tried to extract the root of each word using three very known algorithms which are: khodja stemmer, Nidal et al stemmer, and our proposed stemmer, the obtained results are shown in table 8.

In the other hand, we illustrated the obtained results when applying the three above algorithms on the three corpus used in the experimentation, namely: the small corpus, the middle corpus, and the large corpus, and then we summarized the obtained accuracy for each algorithm in table 9.

Table 8. Extraction of some words roots using the three algorithms

Word	Extracted root			
	Khodja algorithm	Nidal et al algorithm	Our proposed algorithm	Correct root
يتعلّمون	علم	علم	علم	علم
كاتب	كتب	كتب	كتب	كتب
كتاتيب	Not stemmed	كتب	كتب	كتب
اقتصاد	قصد	اقتصد	اقتصد	اقتصد
سنستدرجهم	Not stemmed	درج	درج	درج
متلألئ	Not stemmed	لألأ	لألأ	لألأ
المربّي	ربأ	ربّى	ربّى	ربّى
المربّون	ربن	ربّى	ربّى	ربّى
طائرات	طور	طار	طار	طار
ولولة	ليل	ولول	ولول	ولول
وقيعة	قوع	وقع	وقع	وقع
يزنونهم	زنن	نهب	وزن	وزن
زلازل	Not stemmed	تنازل	زلزل	زلزل
حواسيب	Not stemmed	نسي	حسب	حسب
نوازل	نزل	تنازل	نزل	نزل

Table 9. Illustration of obtained accuracy for the three algorithms

Corpus	Size		The obtained accuracy (suc_ rate, err_ rate)%					
	Nb.roots	Nb.words	Khodja algorithm		Nidal et al algorithm		Our proposed algorithm	
Small	25	50	68,00	32,00	92,00	8,00	98,00	2,00
Middle	135	270	83,70	16,30	63,33	36,66	94,07	5,93
Large	450	1500	73,26	26,74	57,79	42,21	90,53	9,47

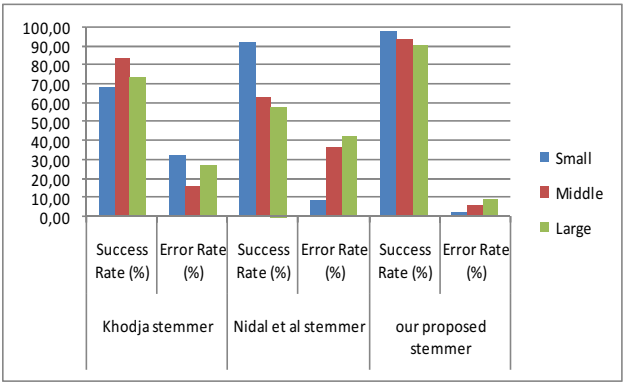


Fig. 3. Comparison between three algorithms

6 Discussion

From table 8, we see that khodja stemmer algorithm fails sometimes in getting the correct root of the given word and for many words it produced one of two results: (1) not stemmed (i.e., حواسيب , متلألئ , سنستدرجهم) completely a new word and sometimes a wrong word that does not exist in Arabic language (i.e.,(ربن ، المربّون) (طور ، طائرات) , (قوع ، وقيعة)). The same thing can be said for Nidal et al algorithm although it's gives best results than khodja algorithm, but it fails for many words like : (زنن ، (سجد، ناسج) (سجد، ناسج) ,(يزنونهم) (نسي، حواسيب). For the same cases, our algorithm gives always the correct root and the failure in our algorithm is very limited.

From Table 9 and figure 3, we can deduce that our proposed algorithm gives the best results for the three used corpus with a very high accuracy. We note here the value 98 % for the small corpus, 94,07 % for the middle corpus, and 90,53 % for the large corpus

7 Conclusion and Perspectives

In this paper we have studied how we can reduce the size of terms in Arabic text categorization by replacing many words by their common root. In this purpose, we exposed the most known algorithms and techniques in the field, Including morphological algorithms mainly based on the use of morphological rules and grammatical patterns of Arabic, and statistical algorithms which are the newest in the field, and require only simple calculations of distances. We also proposed a new statistical algorithm based on bigrams technique. This algorithm is fast and easy to implement on machine, does not require the removal of affixes nor the use of any morphological rules and grammatical patterns, capable to find all types of roots, i.e., trilateral, quadrilateral, quinquelateral, and hexalateral roots. There is no difference between strong roots and vocalic roots in our new algorithm. We also established a comparison between our proposed algorithm and two other algorithms which are very known in the field, namely: Khodja algorithm, Nidal et al algorithm. The first one fails sometimes in getting the correct root of the given word and for many words it produced one of two results: (1) not stemmed word (2) completely a new word and sometimes a wrong word that does not exist in Arabic. The same thing can be said for second one, although it gives best results than the first, but it fails for many words. For the same cases, our new algorithm gives always the correct root, the failure is very limited, and the obtained success ratio of root extraction is very promising.

In our future work, we plan to apply our new algorithm on corpus of Arabic words with big sizes, to improve the obtained success rate, and to apply it in extracting the root of words in other languages such as English and French.

References

1. Fatma, A.H., Keith, E.: Rule-based Approach for Arabic Root Extraction: New Rules to Directly Extract Roots of Arabic Words. Journal of Computing and Information Technology CIT Journal, 57–68 (2014)
2. Ghazzawi, S.: The Arabic Language in the Class Room, 2nd edn. Georgetown University, Washington DC (1992)
3. ETHNOLOGUE, http://www.ethnologue.com/statistics/size (accessed January 16, 2014)
4. Al-Kamar, R.: Computer and arabic language computerizing. Dar Al Kotob Al-Ilmiya, Cairo (2006)
5. Ghwanmeh, S., Kanaan, G., Al-Shalabi, R., Rabab'ah, S.: Enhanced algorithm for extracting the root of Arabic words. In: Proceeding of the 6th International Conference on Computer Graphics, Imaging and Visualization, August 11-14, pp. 388–391. IEEE Xplore Press, Tianjin (2009)
6. Yousef, N., Al-Bidewi, I., Fayoumi, M.: Evaluation of different query expansion techniques and using different similarity measures in Arabic documents. Eur. J. Sci. Res. 43, 156–166 (2010)
7. Wightwick, J., Gaafar, M.: Arabic Verbs and Essentials of Grammar, 2E (Verbs and Essentials of Grammar Series), 2nd edn., p. 160. McGraw-Hill Companies, Inc. (2007) ISBN-10: 0071498052
8. Al-omari, A., Abuata, B., Al-kabi, M.: Building and Benchmarking New Heavy/Light Arabic Stemmer. In: The 4th International conference on Information and Communication systems (ICICS 2013) (2013)
9. Shereen, K., Garside, R.: Stemming Arabic text. Technical report, Computing Department, Lancaster University (1999), http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps (last visited 1999)
10. Momani, M., Faraj, J.: A novel algorithm to extract tri-literal Arabic roots. In: Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications, May 13-16, pp. 309–315. IEEE Xplore Press, Amman (2007)
11. Al shalabi, R.: Pattern-based stemmer for finding Arabic roots. Information Technology Journal 4(1), 38–43 (2005)
12. Hajjar, A.E.S.A., Hajjar, M.: Zreik, K.: A system for evaluation of Arabic root extraction methods. In: Proceeding of 5th International Conference on Internet and Web Applications and Services (ICIW), May 9-15, pp. 506–512. IEEE Xplore Press, Barcelona (2010)
13. Al-Nashashibi, M.Y., Neagu, D., Yaghi, A.A.: An improved root extraction technique for Arabic words. In: Proceeding of 2nd International Conference on Computer Technology and Development (ICCTD), November 2-4, pp. 264–269. IEEE Xplore Press, Cairo (2010)
14. Al-shalabi, R., Kanaan, G., Al-Serhan, H.: New Approach for Extracting Arabic Roots. In: Proceedings of the International ArabConference on Information Technology (ACIT 20003), Alexandria, Egypt, pp. 42–59 (2003)
15. Rehab, D.: Arabic Text Categorization. The International Arab Journal of Information Technology 4(2), 125–131 (2007)
16. Al-Nashashibi, M.Y., Neagu, D.: Ali. A. Y.: Stemming Techniques for Arabic Words: A Comparative Study. In: 2nd International Conference on Computer Technology and development (lCCTD 2010), pp. 270–276 (2010)

17. Kanaan, G., Al-Shalabi, R., Al-Kabi, M.: New Approach for Extracting Quadrilateral Arabic Roots. Abhath Al-Yarmouk, Basic Science and Engineering 14(1), 51–66 (2005)
18. Ghwanmeh, S., Al-Shalabi, R., Kanaan, G., Khanfar, K., Rabab'ah, S.: An Algorithm for extracting the Root of Arabic Words. In: Proceedings of the 5th International Business Information Management Conference (IBIMA), Cairo, Egypt (2005)
19. Mohamad, A., Al-Shalabi, R., Kanaan, G., Al-Nobani, A: Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness. The International Arab Journal of Information Technology, 9(4) (July 2012), (received February 22, 2010) (accepted May 20, 2010)
20. Al-Shalabi, R., Kanaan, G., Ghwanmeh, S.: Stemmer Algorithm for Arabic Words Based on Excessive Letter Locations. In: IEEE Conference (2008)
21. Shereen, K.: Stemming Arabic Text, http://zeus.cs.pacificu.edu/shereen/research.htm
22. Larkey, L., Connell, M.E.: Arabic information retrieval at UMass in TREC 2010. In: Proceedings of TREC 2001, NIST, Gaithersburg (2001)
23. Larkey, S., Ballesteros, L., Margaret, E.: Improving Stemming for Arabic Information Retrieval: Light Stemming and Occurrence Analysis. In: Proc. of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, pp. 275–282 (2002)
24. Larkey, S., Ballesteros, L., Margaret, C.E.: Light Stemming for Arabic Information Retrieval. In: Arabic Computational Morphology. Text, Speech and Language Technology, vol. 38, pp. 221–243 (2007)
25. Sawalha, M., Atwell, E.: Comparative Evaluation of Arabic Language Morphological Analyzers and Stemmers. In: Proceedings of COLING-ACL (2008)
26. Hawas, F.A.: Exploit relations between the word letters and their placement in the word for Arabic root extraction. Comput. Sci. 14, 27–431
27. Hmeidi, I.I., Al-Shalabi, R., Al-Taani, A.T., Najadat, H., Al-Hazaimeh, S.A.: A novel approach to the extraction of roots from Arabic words using bigrams. J. Am. Soc. Inform. Sci. Technol. 61, 583–591 (2010)
28. Boudlal, A., Belahbib, R., Belahbib, A., Mazroui, A.: A markovian approach for Arabic root extraction. Int. Arab J. Inform. Technol. 8, 91–98 (2011)
29. Yousef, N., Aymen, A.E., Ashraf, O., Hayel, K.: An Improved Arabic Word's Roots Extraction Method Using N-gram Technique. Journal of Computer science JSC 10(4) (2014), Published Online http://www.thescipub.com/jcs.toc
30. Frakes, W.B.: Stemming Algorithms. In: Frakes, W.B., Baeza-Yates, R. (eds.) Information Retrieval: Data Structures and Algorithms, pp. 131–160. Prentice-Hall India (1992) ISBN-10: 8131716929

Restoration of Arabic Diacritics Using a Multilevel Statistical Model

Mohamed Seghir Hadj Ameur(✉), Youcef Moulahoum, and Ahmed Guessoum

NLP, Machine Learning and Applications (TALAA) Group
Laboratory for Research in Artificial Intelligence(LRIA)
Department of Computer Science, University of Science and Technology Houari Boumediene (USTHB)
Bab-Ezzouar, Algiers, Algeria
{mohamedhadjameur,moulahoum.youcef}@gmail.com, aguessoum@usthb.dz

Abstract. Arabic texts are generally written without diacritics. This is the case for instance in newspapers, contemporary books, etc., which makes automatic processing of Arabic texts more difficult. When diacritical signs are present, Arabic script provides more information about the meanings of words and their pronunciation. Vocalization of Arabic texts is a complex task which may involve morphological, syntactic and semantic text processing.
In this paper, we present a new approach to restore Arabic diacritics using a statistical language model and dynamic programming. Our system is based on two models: a bi-gram-based model which is first used for vocalization and a 4-gram character-based model which is then used to handle the words that remain non vocalized (OOV words). Moreover, smoothing methods are used in order to handle the problem of unseen words. The optimal vocalized word sequence is selected using the Viterbi algorithm from Dynamic Programming.
Our approach represents an important contribution to the improvement of the performance of automatic Arabic vocalization. We have compared our results with some of the most efficient up-to-date vocalization systems; the experimental results show the high quality of our approach.

Keywords: Statistical language model · Arabic language · Hidden markov model · Automatic vocalization · Dynamic programming · Smoothing · Corpus · Viterbi algorithm.

1 Introduction

Arabic texts are generally written without diacritical signs (newspapers, books, etc.) this does not pose a problem for people who have a certain mastery of Arabic since they can easily infer the diacritical signs from the context of the words. However, this can be problematical for non-native Arabic speakers. As a matter of fact, the absence of diacritical signs in words also makes their automatic processing more difficult. Indeed, when diacritics are present, the Arabic

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 181–192, 2015.
DOI: 10.1007/978-3-319-19578-0_15

script provides more information about words meanings and their pronunciations. As such, Arabic vocalization is used in order to increase the performance of many applications such as Arabic text-to-speech (TTS) [1,10] and speech recognition [16].

Arabic diacritics restoration (text vocalization) is the process of assigning Arabic diacritics such as fatha ("a" sound as in "apple"), damma ("oo" sound as in "book") and kasra ("i" sound as in "in") to a given text (or script). Arabic diacritical signs are represented in Table 1.

Table 1. Arabic diacritical signs

Diacritic	Example	Pronunciation
Fatha	ذهبَ	/t//a/
Damma	الطفلُ	/t//u/
Kasra	البيتِ	/t//i/
Tanween Damma	كِتابٌ	/t//un/
Tanween Kasra	كِتابٍ	/t//in/
Tanween Fatha	كِتابًا	/t//an/
Sukuun	الوقْتُ	/t/
Shadda	مُدَرِّسَة	/t//t/

During the last few years, the statistical approaches have been proven to be more efficient in the tackling of different problems of natural language processing. For the vocalization problem more specifically, most of the recent work was based on statistical approaches which can be either purely statistical ones or hybrid methods that combine a statistical language model and some other treatments.

Hybrid methods, such as [7] which uses a morphological tagger or [4] which uses *AlKhalil Morpho Sys* [4], depend on the effectiveness (accuracy) of these morphological analysers and taggers. Purely statistical methods however do not have such a dependence. Recent works based on purely statistical methods have reported very interesting results. This is the case for [9] which uses only a word-based bigram language model and the work of [2] which uses a character-based 4-gram model.

In this paper, we aim to further improve the previous statistical Arabic text vocalization approaches used in [9] and [2] by proposing a new simple but efficient system that relies on a purely statistical language model coupled with dynamic programming which combines these two approaches; thus our vocalization system is based on two models: the first one is a bi-gram word-based model [9] which is first used for vocalization and the second one is a 4-gram character-based model [2] which is used to handle the words that remain non-vocalized (OOV words). Smoothing methods are used in order to handle the problem of unseen words; the optimal vocalized word sequence is selected using the Viterbi algorithm [12].

This paper is organized as follows: Section 2 gives an overview of the state of the art vocalization systems. Section 3 explains our approach to restoring Arabic diacritics using a statistical language model and dynamic programming. Section 4 presents our tests and experimental results. A conclusion of our work is given in Section 5.

2 Related Work

Vocalization approaches can be divided into two main categories: Rule-based and Statistical Approaches. During the last decade, the statistical approaches have widely been used in a variety of natural language processing applications which have proven their efficiency. For the vocalization problem, most of the recent work was based on statistical approaches. These statistical approaches can be classified into two categories: purely statistical methods, or hybrid methods that combine a statistical language model and some other treatments.

In terms of purely statistical methods, one may cite [6] where the authors presented a vocalization approach based on Hidden Markov Models (HMMs). The hidden states correspond to the vocalized words and each one of them has a single emission leading to a non vocalized word (an observed state). In [2], a similar approach was used but with a character-based 4-gram model (a sequence of 4 consecutive vocalized letters) instead of a word-based model. The most recent work based on purely statistical methods is [9] where its authors used a statistical bigram language model coupled with dynamic programming to choose the most likely sequence of diacrtics. They improved their own work in [8] by using a higher order n-gram statistical language model.

For the methods which use a hybrid approach, we can mention [7] whose authors developed a hybrid system which combines a statistical n-gram language model (where n equals 1, 2 or 3) combined with a morphological tagger. In a similar way, in [3] a statistical n-gram language model is also used along with morphological analysis using *AlKhalil Morpho Sys* [4]. In [15], an approach was proposed which combines lexical retrieval, bigram-based and SVM-statistical prioritized techniques. In [14], the authors proposed two methods: the first uses an n-gram statistical language model along with A^* lattice search while the second method attempts to segment each Arabic word into all its possible morphological constituents then proceed in a similar way as the first one. The authors reported that their second approach gives better results. Finally, in [17], a statistical classifier was proposed which is based on the maximum entropy principle, which uses the combination of a wide array of lexical, segment-based and part-of-speech tag features in order to select the best classification.

It turns out that Arabic text vocalization is not yet optimal as will be shown in Section 4.4. No system is currently good enough to restore diacritics with high enough a quality as to be able to build solid applications on it. For this reason, we have decided to dig deeper into this problem. This has led us to building a system which has given very encouraging results as will be shown in the sequel.

3 Arabic Text Vocalization Approach

In this section we will formally introduce the problem of Arabic text vocalization and present the different models generated in our system.

3.1 Formalizing the Problem

Vocalization of Arabic text (or Restoration of Arabic Diacritics) is the process of assigning diacritical signs to each word in a given text or script.
This problem can be formalized as follows: given a sequence of non-vocalized words (or script) $W = w_1, w_2, ..., w_n$, the vocalization task is to find the best sequence of vocalized words $V = v_1, v_2, ..., v_n$ from all the possible vocalization sequences of W.

Assigning a score to each possible vocalized word sequence can be used to select the best vocalization from all the possible ones. This score can be calculated using the Chain rule:

$$P(W) = \Pi_{k=1}^{n} P(W_k | W_1^{k-1}) \tag{1}$$

By making the independence assumption (Markov assumption) for the n-grams model, instead of using the whole history (chain-rule) the n-gram model can approximate the history of a given word using just the last k words. The probability will thus be estimated as follows:

$$P(W_n | W_1^{n-1}) = P(W_n | W_{k-(n-1)}^{n-1}) \tag{2}$$

Using the Markov assumption in the case of a bi-gram language model, we will have:

$$P(W_n | W_1^{n-1}) = P(W_n | W_{n-1}) \tag{3}$$

$P(W_n|W_{n-1})$ is computed using the Maximum Likelihood Estimation (MLE):

$$P(W_n | W_{n-1}) = \frac{C(W_{i-1}, W_i)}{C(W_{i-1})} \tag{4}$$

where $C(W_{i-1}, W_i)$ and $C(W_{i-1})$ are the counts of the bi-gram $W_{i-1}W_i$ and the uni-gram W_{i-1} respectively.

3.2 Presentation of the Vocalization System

This section presents the global structure of our vocalization system. The automatic vocalization of Arabic texts consists of two main phases: vocalization using a bi-gram word-based model followed, for the unresolved cases, by vocalization using a 4-gram character-based model. This is illustrated in Figure 1 and explained in more details in the following two subsections. We should point out that we have decided for our word-based model to restrict ourselves to bi-grams for computational efficiency reasons. Going for higher-order n-Gram models would indeed be costly in execution time in an application which should be as fast as possible to be integrate into larger, possibly online, applications. As to the 4-gram character-based model, this is due to the fact that we have analyzed that 4 letters are can be quite rich a background to allow for reasonably good diacritization, especially that this second model is used as a complement to the word-based one.

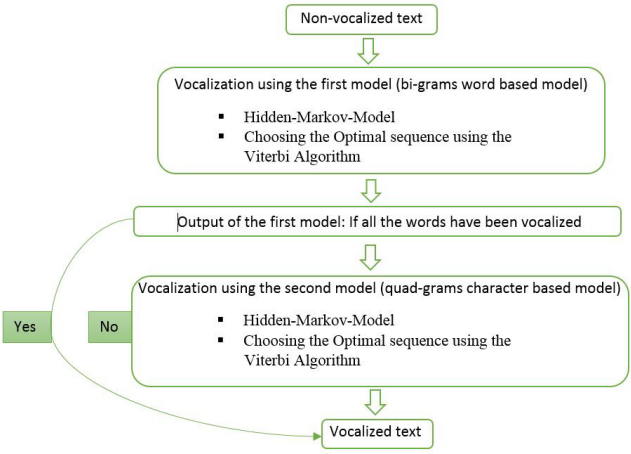


Fig. 1. The Vocalization System

3.3 The First Model

Our first model is a bi-gram-based language model. To illustrate it, let us consider the following non-vocalized sentence:

خلق الإنسان علمه البيان

The functioning of the first model can be summarized in the following steps:

The first step is to build a dictionary which associates for each non vocalized word all its possible vocalizations.

In the second step, a lattice is created for the non-vocalized sequence $W = w_1, w_2, ..., w_n$ which gives, for each non-vocalized word w_i, all its possible vocalizations from the dictionary. In order to simplify the explanation, our example considers only a subset of the possible vocalizations of each word (as shown in Figure 2). Given the assumption, the size of the subset of possible vocalizations would be $8 * 4 * 8 * 2 = 512$.

The third step consists in associating to each possible vocalization (e.g. in Figure 2, each sentence among the 512 possible ones) a probability using the bi-gram language model:

$$P(W_n|W_1^{n-1}) = P(W_n|W_{n-1}) \tag{5}$$

For example one of the possible vocalizations is:

خَلَقَ الْإِنْسَانَ عَلَّمَهُ الْبَيَانَ

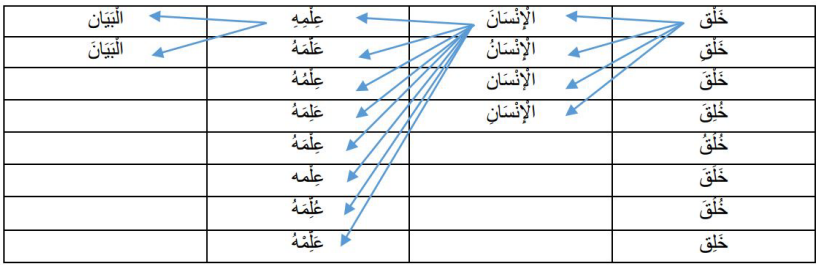


Fig. 2. Some of the possible vocalizations for a non-vocalized sentence

The probability of the above sentence is calculated as follows:

$$P\ (\text{خَلَقَ الْإِنْسَانَ عَلَّمَهُ الْبَيَانَ}) =$$

$$P\ (\text{خَلَقَ}) * P\ (\text{الْإِنْسَانَ} \mid \text{خَلَقَ}) * P\ (\text{عَلَّمَهُ} \mid \text{الْإِنْسَانَ}) * P\ (\text{الْبَيَانَ} \mid \text{عَلَّمَهُ})$$

Similarly, probabilities are assigned to all the possible vocalized word sequences.

The forth and final step is to find among all the possible vocalizations the one that has the highest probability.

$$v_1, \widehat{v_2, ...}, v_n = argmax(\Pi_{k=1}^{n} P(v_k|v_{k-1})) \tag{6}$$

The number of all possible vocalizations is very large as mentioned in Figure 2. Let N be the average number of all the possible vocalizations for each word and L the length of the non vocalized sequence (i.e. the number of words in it). The number of all possible vocalizations will then be N^L. Trying to find the best vocalization by a brute-force approach would have an exponential complexity ($O(N^L)$) and is clearly not efficient. An alternative is to use , the Viterbi algorithm, a Dynamic programming approach. To this end, let us present our Hidden Markov Model (HMM) which will be used as an input to the Viterbi algorithm to get the best vocalization.

Hidden Markov Model Our Hidden Markov Model (HMM) is defined by:

- A set of states which represent the vocalized words $v_1, v_2, ..., v_n$.
- A set of observations which represent the non-vocalized words $w_1, w_2, ..., w_n$.
- The transition matrix which contains the transition probabilities $P(v_i|v_{i-1})$

Generally each sequence in the HMM depends on two probabilities (transitions and emissions). In our model however, only the transition probabilities are considered.

The Viterbi Algorithm. The best vocalized sequence is chosen from the HMM using the Viterbi algorithm which is a very efficient algorithm for selecting the best vocalization sequence [12]. The latter uses a recursive relation in which the probability of each node at a given level i is calculated according to its preceding level $i-1$ (see Figure 3).

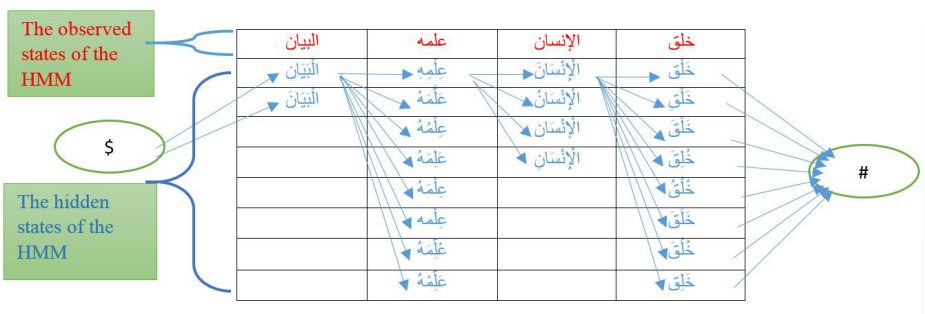


Fig. 3. Finding the optimal vocalized sequence using the Viterbi algorithm

As shown in Figure 3, the weight of each node of index (i, j) of level i is calculated from all the nodes of its preceding level $i-1$, using the following formula:

$$P(i,j) = max_{k=1,v_{i-1}}(P(i,j|i-1,k) * p(i-1,k)) \quad (7)$$

where i and j are the indexes of the line and colon, respectively, in the transition matrix.
v_{i-1} is the number of all possible vocalization for word $i-1$.
After calculating the weights of all nodes (in this forward-moving computation), while keeping track of the best nodes on this path, back-tracing is done in order to find the optimal path.

The Viterbi algorithm allows to efficiently solve the problem of selecting the best vocalization sequence.

In the next section, we introduce the smoothing method we use in order to handle the problem of unseen bigrams.

Handling the Problem of Unseen bi-grams. The maximum likelihood estimation (MLE) is calculated using Equation 4. This equation assigns a null probability to any bi-gram that does not belong to the training corpus. According to Jurafsky and Martin ([11]), almost 99% of all the possible combinations of bi-grams may be missing from any given corpus. This is why the use of smoothing methods are necessary in order to avoid having null probabilities (as a result of the products of probabilities). This problem is handled by taking some of the probability mass from the existing n-grams and distributing it to the non-found n-grams.

Additive Smoothing: One of the simplest smoothing methods is Additive Smoothing [5] which assumes that each of the n-grams occurs one more time than its actual occurrence count. Thus we add one to all the n-grams. This yields in the case of bi-grams:

$$P_{add}(w_i|w_j) = \frac{K + c(w_i, w_j)}{(K * V) + c(w_i)} \tag{8}$$

where K is a constant between 0 and 1 and V is the size of the vocabulary.

Absolute Discounting: Absolute discounting [11] is an interpolated smoothing method [13] which is obtained by discounting a constant D between 0 and 1 from each non-null probability mass. This yields in the case of bi-grams:

$$P_{abs}(w_i|w_j) = \frac{max(c(w_i, w_j) - D, 0)}{c(w_i)} + \frac{D}{c(w_i)} N_{1+}({w_i}^*) P_{abs}(w_j) \tag{9}$$

where $P_{abs}(w_j) = \frac{1}{V}$, V is the vocabulary size.
$N_{1+}({w_i}^*)$ is the number of all the words without repetition that follow w_i in the training corpus. In the next section we explain our second model.

3.4 Letter-Based Model

The second model is a 4-gram character-based model which is used to handle the words that remain non-vocalized (out of vocabulary words, OOV). The HMM used in this model is now introduced.

Letter-Based Hidden Markov Model Our HMM is defined by states that represent the vocalized letters and observations that represent the non-vocalized letters. It consists of:

- A set of states which represent the vocalized letters $q_1, q_2, ..., q_n$.
- A set of observations which represent the non-vocalized letters $l_1, l_2, ..., l_n$.
- The transition matrix which contains the transitions $P(q_i|q_{i-1}, q_{i-2}, q_{i-3})$.
- The emission matrix which contains $P(l_i|q_i)$.

This model is used in a similar way to the previous model. The same smoothing methods are used and the optimal path is selected using the Viterbi algorithm. This model has the capacity to vocalize any Arabic word; this is why it is used as a final step to ensure the complete vocalization of the non-vocalized script.

4 Implementation and Tests

In this section, we start by presenting the corpus we have used and some statistics relating to it. We then move to presenting the results of testing our implementation.

The source code of our vocalization system is available at https://github.com/Ycfx/Arabic-Diacritizer under the GNU General Public License (*GPL*).

4.1 Corpus Construction

The largest part of our corpus is automatically retrieved from the site http://www.al-islam.com/ using a URL-rule-based crawler. This site is an Islamic religious site that contains vocalized text about a number of subjects (Hadith, Commentaries of the Quran, etc.). A vocalized *Holy Quran* was also downloaded from http://tanzil.net/ and added to the corpus. Each downloaded vocalized text goes through cleaning, tokenisation, and normalisation steps to finally yield a properly vocalized corpus. On the other hand, its non-vocalized version is obtained by simply deleting the diacritical signs from the vocalized corpus.

4.2 Corpus Statistics

We have created a large Arabic corpus which contains more than 10 million words (tokens) 2. We have used 90% of the total words of the corpus for the training phase and the remaining 10% for the testing phase.

Table 2. Corpus statistics

	Corpus
Sentences	799 470
Tokens	10 634 921
Types	379 429

4.3 Evaluation Measures

To measure the performance of the different vocalization systems we have used the Word Error Rate (*WER*) and the Diacritic Error Rate (*DER*) measures:

- *WER1*: the number of words vocalized wrongly by the system (taking into account the diacritic of the last letter of the word).
- *WER2*: the number of words vocalized wrongly by the system (not taking into account the diacritic of the last letter of the word).
- *DER1*: the number of characters vocalized wrongly by the system (taking into account the diacritic of the last letter).
- *DER2*: the number of characters vocalized wrongly by the system (not taking into account the diacritic of the last letter).

4.4 Results

In this section we will start by presenting the results obtained by changing the smoothing parameters, then we give a detailed comparison of our system with the state of the art vocalization systems.

Table 3. The impact of the smoothing parameters on the vocalization system

Smoothing methods	WER1	WER2	DER1	DER2
Absolute discounting (K=1)	11.85	6.67	4.63	3.49
Absolute discounting (K=0.5)	11.57	6.30	4.34	3.23
Absolute discounting (K=0.1)	**11.53**	**6.28**	**4.30**	**3.18**
Additive smoothing (D=1)	16.87	9.49	8.10	6.86
Additive smoothing (D=0.5)	15.75	9.16	7.85	6.83
Additive smoothing (D=0.1)	15.41	9.05	7.77	6.83

Impact of the Smoothing Parameters on the Vocalization System. Table 3 shows the impact of changing the smoothing parameters on the vocalization system. The K and D smoothing parameters were investigated for additive and absolute discounting methods respectively.

The experimental results prove that the use of smoothing methods has a noticeable influence on the overall performance of the vocalization system. The best performance of our system was achieved using absolute discounting with $D = 0.1$, where the results we obtained were 11.53% in terms of Word Error Rate *(WER1)* and 6.28% when the case ending was ignored (*WER2*), and in terms of Diacritic Error Rate the results were 4.30% for *DER1* and 3.18% when ignoring the last diacritical mark (*DER2*). These results show that the Absolute Discounting method gives a better practical performance in comparison to Additive Smoothing.

Comparison of our System with the Different Vocalization Systems. In order to evaluate the overall performance of our vocalization system, we have compared its performance to some of the most efficient implementations available today. However since these systems have not been tested on the same corpus, the conclusions should be taken with some caution. The results of the comparison are summarized in Table 4.

When case ending (the last diacritical sign) is ignored *(WER2,DER2)*, the results were clearly better for all the compared systems, which is explained by the added difficulty when attempting to vocalize the last letter *(WER1,DER1)*.

Table 4. Comparing the performance of our system to those of some other vocalization systems

Vocalization Systems	WER1	WER2	DER1	DER2
Zitouni et al (2006) [17]	37	15	23	10
Habash et al (2007) [7]	14.9	4.8	5.5	2.2
Shaalan et al (2009) [15]	12.16	**3.78**	-	-
Rashwan et al (2011) [14]	12.5	3.8	**3.1**	**1.2**
Hifny et al 2013) [9]	12.5	7.4	-	-
Bebah et al (2014) [3]	21.11	9.93	7.37	3.75
Our system	**11.53**	6.28	4.30	3.18

Our system gives the best result in terms of *WER1*, next to best on *DER1*, and its performance on the other measures is very close to the best results reported in the literature. That our system performs extremely well on *WER1* and *DER1* is indeed what we want; it shows that other systems have problems handling the last letter diacritic which is very crucial in Arabic.
Our system performance has thus been proven to be very competitive; It shows the effectiveness of the multilevel approach we have adopted for the vocalization problem.

5 Conclusion

We have presented in this paper a multilevel statistical vocalization model. Our system is based on two models: the first one is a bi-gram word-based model which is used first for vocalization and the second one is a 4-gram letter-based model which is used as a back-off, i.e. to handle the words that remain non-vocalized after the application of the first model. We have used smoothing methods to handle the problem of unseen words and the Viterbi algorithm to select the optimal path, i.e. best vocalization sequence, in the HMM. The results shows the efficiency of our vocalization system in comparison to other state-of-the-art systems.

Our system can be improved in several ways which we intend to explore:

- Our first model is based on bi-gram probabilities only; the use of n-gram models with $n > 2$ should yield better results.
- We can enrich the corpus by adding many more modern Arabic texts to it.
- In this work, only two smoothing methods have been used; using other smoothing methods could give better results.

References

1. Ahmed, M.E.: Toward an arabic text to speech system. The Arabian Journal of Science and Engineering 16(4B) (1991)
2. Alghamdi, M., Muzaffar, Z.: Kacst arabic diacritizer. In: The First International Symposium on Computers and Arabic Language, pp. 25–28 (2007)
3. Bebah, M., Amine, C., Azzeddine, M., Abdelhak, L.: Hybrid approaches for automatic vowelization of arabic texts. arXiv preprint arXiv:1410.2646 (2014)
4. Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M., Shoul, M.: Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In: International Arab Conference on Information Technology (2010)
5. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics, pp. 310–318. Association for Computational Linguistics (1996)
6. Gal, Y.: An hmm approach to vowel restoration in arabic and hebrew. In: Proceedings of the ACL 2002 Workshop on Computational Approaches to Semitic Languages, pp. 1–7. Association for Computational Linguistics (2002)

7. Habash, N., Rambow, O.: Arabic diacritization through full morphological tagging. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, pp. 53–56. Association for Computational Linguistics (2007)
8. Hifny, Y.: Higher order n gram language models for arabic diacritics restoration. In: Proceedings of the 12th Conference on Language Engineering, Cairo, Egypt (2012)
9. Hifny, Y.: Restoration of arabic diacritics using dynamic programming. In: 2013 8th International Conference on Computer Engineering & Systems (ICCES), pp. 3–8. IEEE (2013)
10. Hifny, Y., Qurany, S., Hamid, S., Rashwan, M., Atiyya, M., Ragheb, A., Khallaaf, G.: An implementation for arabic text to speech system. In: The Proceedings of the 4th Conference on Language Engineering (2004)
11. Jurafsky, D., Martin, J.H.: Speech and language processing. Pearson Education India (2000)
12. Neuhoff, D.L.: The viterbi algorithm as an aid in text recognition. IEEE Transactions on Information Theory 21(2), 222–226 (1975)
13. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependences in stochastic language modelling. Computer Speech & Language 8(1), 1–38 (1994)
14. Rashwan, M.A., Al-Badrashiny, M.A.S., Attia, M., Abdou, S.M., Rafea, A.: A stochastic arabic diacritizer based on a hybrid of factorized and unfactorized textual features. IEEE Transactions on Audio, Speech, and Language Processing 19(1), 166–175 (2011)
15. Shaalan, K., Abo Bakr, H.M., Ziedan, I.: A hybrid approach for building arabic diacritizer. In: Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pp. 27–35. Association for Computational Linguistics (2009)
16. Vergyri, D., Kirchhoff, K.: Automatic diacritization of arabic for acoustic modeling in speech recognition. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, pp. 66–73. Association for Computational Linguistics (2004)
17. Zitouni, I., Sorensen, J.S., Sarikaya, R.: Maximum entropy based restoration of arabic diacritics. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 577–584. Association for Computational Linguistics (2006)

A New Multi-layered Approach for Automatic Text Summaries Mono-Document Based on Social Spiders

Mohamed Amine Boudia[1(✉)], Reda Mohamed Hamou[2], Abdelmalek Amine[3], Mohamed Elhadi Rahmani[4], and Amine Rahmani[5]

Dr. Moulay Tahar University SAÏDA
Department of Computer Saida, Laboratory Knowledge Management
and Complex Data (GeCoDe Lab), Dr. Moulay Tahar University SAÏDA, Saida, Algeria
{mamiamounti,hamoureda,abd_amine1}@yahoo.fr,
r_m_elhadi@yahoo.fr, aminerahmani2091@gmail.com

Abstract. In this paper, we propose a new multi layer approach for automatic text summarization by extraction where the first layer constitute to use two techniques of extraction: scoring of phrases, and similarity that aims to eliminate redundant phrases without losing the theme of the text. While the second layer aims to optimize the results of the previous layer by the metaheuristic based on social spiders. the objective function of the optimization is to maximize the sum of similarity between phrases of the candidate summary in order to keep the theme of the text, minimize the sum of scores in order to increase the summarization rate, this optimization also will give a candidate's summary where the order of the phrases changes compared to the original text.The third and final layer aims to choose the best summary from the candidate summaries generated by layer optimization, we opted for the technique of voting with a simple majority.

Keywords: Automatic summary extraction · Data mining · Social spider · Optimization · Scoring similarity

1 Introduction and Problematic

Every day, the mass of electronic textual information is increasing, making it more and more difficult access to relevant information without using specific tools. In other words access to the content of the texts by rapid and effective ways is becoming a necessity.

A summary of a text is an effective way to represent its contents, and allow quick access to their semantic content. The purpose of a summarization is to produce an abridged text covering most of the content from the source text.

« We can not imagine our daily life, one day without summaries », underline Inderjeet Mani. Newspaper headlines, the first paragraph of a newspaper article, newsletters, weather, tables of results of sports competitions and library catalogs are all summarized. Even in the research, the authors of scientific articles must accompany their scientific articles by a summary written by themselves.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 193–204, 2015.
DOI: 10.1007/978-3-319-19578-0_16

Automatic summary can be used to reduce the search time to find the relevant documents or to reduce the treatment of long texts by identifying the key information.

Our work uses automatic summarization by extraction, because it is a simple method to implement and gives good results; only in the previous works, produce the automatic summary by extraction consists to use only one technique at a time (Score, Similarity sentence or prototype) and respects the order of the sentences in the original document, our work answers the following questions:

- What is the contribution of the use of two methods of summarization at the same time on the quality of summary?
- Can the bio-inspired method based on the social spiders brings more for the automatic summary and increase the quality of the summary?

2 Our Proposed Approach

To create a summary by extraction, it is necessary to identify textual units (phrases, clauses, sentences, paragraphs) considered salient (relevant), then the select the textual units that hold the main ideas of the text with a certain order, in order to build a summary.

The approach presented in this article obeys the following steps:

2.1 Pretreatment

Simple cleaning: a stop words will not be removed, because the method of automatic summarization by extraction aims to extract the most informative sentences without modifying them: if we remove the empty words without information on their morpho-syntactic impact in sentences, we risk having an inconsistent summary of a morphological point of view.

Then cleaning is to remove emoticons, to replace spaces with "_" and remove special characters (#, \, [,] .....).

– Choice of term: for automatic summarization by extraction we will need two representations:

- Bag of words representation
- Bag of sentence representation.

Both representations are introduced in the vector model.

The first representation is to transform the text into a vector v_i (w_1, w_2, ..., $w_{|T|}$) where T is the number of all the words that appear at least once in the text. The weight w_k indicates the occurrence of t_k word in the document.

The second representation is to transform the text into a V'I vector (q_1, q_2, ..., $q_{|R|}$) where R is the number of all the phrases that appear at least once in the text. The q_k weight indicates the occurrence of t_k sentence in the document.

And finally a word phrases- occurrence matrix will be generated after the two previous representation, the size of this matrix is equal to (the number of words in the text) X (the number of words in the text); p_{ik} weight is the number occurrence of the word *i* in the sentence *j*;

2.2 Layer 1 : Pre-summary

Weighting and Pre-summary.

Weighting.
Once the "Word-Phrase" matrix is ready, we calculate a weighting of "Word-Phrase" matrix using a known encodings (tf-idf or tfc) with a small modification to the adapted the concept of a mono-document summarization.

The weight of a term in a sentence t_k p_i is calculated as:

- *TF-IDF.*

$$tf - idf(t_k, p_i) = tf(t_k, p_i) * \log(\frac{A}{B}) \quad (1)$$

$tf(t_k, p_i)$: the number of occurrences of the term tk in the phrase pi;
A : the total number of sentences in the text;
B : the number of sentences in which the tk term appears at least once.

- *TFC.*

$$\text{tfc}(\text{t}_\text{k}\ , \text{p}_\text{i}) = \frac{\text{tf-idf}\ (\text{t}_\text{k}\ , \text{p}_\text{i})}{\sqrt{\sum_{\text{i=1}}^{|\text{p}|}\ \text{tf-idf}\ (\text{t}_\text{k}\ , \text{p}_\text{i})^2}} \quad (2)$$

After calculating the weighting of each word, a weight is assigned to each sentence. The generated summary is then generated by displaying the highest score of the source document sentences.

This score of a sentence is equal to the sum of the words in this sentence:

$$\text{SCORE}\ (\text{p}_\text{i})\ = \sum_{k=0}^{nbr\_word} Mik \quad (3)$$

Primitive summary
"Suggested process claims on the principle that high-frequency words in a document are important words" [Luhn 1958]

The final step is to select the N first sentences that have the highest weight and which are considered the most relevant. The process of extracting the first N sentences intended to build the summary is defined either by a threshold, in this case, the score of the sentence must be greater than or equal to the threshold in order that this sentence will be extracted the second method is to fix a number N of phrase to be extracted, all phases will be ranked in descending order according of their score, and we take only the first N phrases.

Elimination of Rehearsals and Theme Detection: Using SIMILARITY Method Summarization by Extraction
The result of the previous step is a set of phrases which is a high score. Just we have a possibility that two or more sentences have a high score but they are similar, so we proceed to the elimination of phrases that resembling. The similarity between the

sentences that have been selected at the end of the previous step with known metrics (Euclidean .....).

Two parameters are used to adjust the elimination of repetitions: similarity threshold and reduction rate, the first parameter defined the point that we can consider two sentences as similar, and the second parameter indicates the number of resemblance to eliminate, to decrease the entropy information. When the similarity between two sentences is greater: they the phrase that has the highest score stay and we remove the other sentence.

The similarity is also used to detect the sentence that has more relation with the them of the text. According to the domain experts, it is the sentence which is most similar to the other sentences holds the theme text.

2.3 Layer 2: Optimization Using Social Spiders

Optimization Using Social Spiders

Natural Model

- **Environment**: a set of pickets which serve weaving wire brackets, this pickets have different sizes.
- **Weaving**: weaving is to create a link between the current and the last visited pick
- **Movement**: movement allows the spider to move in the environment on the wire woven by her or by others spider in the same canvas. The selection of the new position dependent upon a finite number of criteria. The wire has a flexibility F which is one of the major criteria of movement of the spider, the flexibility F represents the maximum weight with a correction relative to its diameter that can be held by the wire.
- **Communication**: social spiders communicate with the others in the weaving task, movement or capture prey; communication can be done by two different methods; by vibration on the wire or by the concentration of hormonal substances that spider left on a wire. Each vibration intensity and each concentration of substances has a specific meaning to others spiders, and this means that each spider must have two receivers (vibration and concentration).
- **System dynamics**: It is built on the principle of stigmergy: the behavior of agents have effects on the environment (wire-laying) in return behavior is influenced by the context of the agent (environment).

Artificial Model

- **Environment :** a picket grid (N * N). N is the square root of the number of phrases after layer 1 (pre-summary and the elimination of similar phrases) where each pickets is representing a sentence, the pickets have different sizes that representing the score of the phrase. Initially, all the wires are woven so as to have a complete graph;

The number of spiders is equal to or less than the number of phrases, each spider is placed on a pole (phrase) randomly.

- **Weaving :** the wires are woven in the beginning of each iteration in order to have a complete graph. The similarity s_{xy} between two phrases ph_x ph_y represents the diameter of wire woven between the two picket x and y associate to phrases ph_x to and p_{hy}, as given the similarity is commutative (s (x, y) = s (y, x)): the diameter of wire woven between the two picket x and y will be a uniform.
- **Movement:** movement of the spider is incremental and random; Every spider save in its memory every way which she followed. To save result, a weight of path should be: Superior to the "lower threshold of the summary rate" and Less the "upper threshold of the summary of rates."

  We associated to the social spider *i* in iteration *j* a P_{ij} weight initialized to zero and it equal to the sum of the weights of *k* SCORE sentences whose social spider *i* have visited during the iteration *j.*

  The wire has a flexibility that F depends on its diameter is constant and represents the maximum weight that can load on itself, with artificial model F Is defined as follows.

$$Flexible\left(fil_{ij}\right) = Seuil\ supérieur\ de\ taux\ de\ résumé\ *\ \ diametre\left(fil_{ij}\right) \quad (4)$$

$$diametre\left(fil_{ij}\right) = \ similarité\left(\ phrase_i, phrase_j\right) \quad (5)$$

Noting that:

- Abstract rate threshold is constant.

If i social spider during operation j with Pij weight passes through the wire (x, y), it will execute this pseudo-algorithm:

```
If P_ij is lower than F(x,y) then
  the spider will go to the wire (x,y)
  updating the current path
  Update the weight P_ij,
If the wire is torn
```

Social spider i will go into pause waiting for the end of the iteration j.

We will give these two observations:

(a) F(x,y) is higher than F(w,z) is equivalent to say that the similarity between the sentence x and the sentence y is greater than the similarity between the sentence w and the sentence z because “upper threshold of the summary of rates” is constant.

(b) The interpretation of F(x,y) is higher than F(w,z), is that by optimizing with social spider: if choice between wire (x,y) and the wire (w,z) the spider will choose the first wire because it safe for her . If his current weight Pij is high; the second wire risk to tear.

From observations A and B, we can deduce that the optimization is to minimize the weight of the summary, to maximize the similarity to preserve the theme of the candidate summary, while respecting the dice constrained utility and semantics represented

by the interval [lower threshold summary of rates, upper rates higher threshold] noting that the lower and upper thresholds are summarized determined and fixed as language experts.

- ***The utility constraint*:** Automatically produce a summary with higher summary score "upper threshold of the summary of rates," is not helpful.
- ***The semantic constraint:*** Automatically produce a summary with lower summary score "lower threshold of the summary of rates," losing a lot of semantics.

— **End of iteration :** when all the spider will be in pause state, the iteration j will be declared finished, the spider will rewoven the spiders randomly choose their new start position and start the iteration j + 1.

— **Communication :** Each spider leaves a trace on hormonal stakes visited so that other spiders will not take this part of the way. First it ensures diversity between different summaries candidate that is greater coverage suspected combination spiders consider this shift pickets the number they share with each spider that operates on the canvas, and moves with the constraint of not exceeding M common stake in the same order with another spider .

Secondly, communication is used to avoid the repetition of sentences in the summary. In cases where social spider returns while moving on a picket that it been already have been visited by itself in current iteration it makes a flashback and continues his trip without considering this visit.

The duration of evaporation of communication hormone spider is equal to an iteration, it should be noted that the hormone density can not be cumulative.

— **System dynamics**: It is built on the principle of stigmergy: the behavior of agents have effects on the environment, in return behavior is influenced by the context of the agent.

Each spider keeps in mind, the best visited paths, after a number of spider iterations, every spider returns the best paths.

— **Path :** is a series of picket visited in chronological order, and is a summarization. Recall that each picket is a phrases (see the initial state).

— **End of the optimisation of the social spiders:** when the number of iterations performed reached the maximum number of iterations, each spider returns all paths (where each path, is a candidate summary). Was associated with each path or summarization ie a set of candidate evaluations indices. And launching a voting algorithm compared these evaluation indices to choose the best candidate summary to remember.

2.4 Layer 3: Evaluation and Vote

Candidates generated by the previous layer abstracts will be evaluated by several evaluations metric, and then we will classify pairs. R1 and R2 are two abstracts candidate rate by N metric evaluation, the number of associated point R1 represents the number of evaluation indicating that the quality of R1 is greater than or equal to R2

and aims to it. The summary with most points will win the duel and will face another until there will be more challenger. Summary will be declared the winner as the best back to resume.

3 Experimentation

Under the assumption that the weight of a sentence indicates its importance in the document and under assumption that two similar sentences have the same meaning; we applied the algorithms summarized by extracting in occurrence Scoring and similarity phrases. Our method is oriented for the moment to the generation of a mono-document summary using a biomimetic approach (Social Spider).

3.1 Used Corpus

Was used as the text corpus "Hurricane" in French, which contains a title and 20 sentences and 313 words, after the pretreatment process and vectorization bag of words, we get 171 different token. And we have took three references summaries produced successively by Summarizer CORTEX, Essential Summarizer, and a summary produced by a human expert.

3.2 Validation

We evaluated the summaries produced by this algorithm with the metric ROUGE (Lin 2004) which compares a candidate summary (automatically produced) and Summary Reference (created by human experts or other automatic summarization systems known).

The Evaluation Measure Recall - Oriented Understudy for Gisting Evaluation
We evaluate the results of this work by the measure called Recall - Oriented Understudy for Gisting Evaluation (ROUGE) proposed by (Lin, 2004) involving the differences between distributions of words.

$$ROUGE\ (N) = \frac{\sum_{s \in R_{ref}} \sum_{s \in R_{can}} \mathrm{Co} - \mathrm{occurences}\ (R_{ref}, R_{can}, N)}{\mathrm{Nbr} - \mathrm{NGramme}\ (N)_{R_{ref}}} \quad (6)$$

F-Measure for the Evaluation of Automatic Extraction Summaries
We have proposed in our work before an adaptation of the F-measure for the validation of automatic summarization by extraction, as this technique is based on phrases to keep and delete

Confusion matrix	Candidate sumarry

Word K : number of words to keep
Word R: number of words to remove

Automatic summary		word K	word R
Reference summary	Word K	X	Y
	Word R	Z	W

Table 1. Adaptation of the F-measure for the validation of automatic summarization

From the confusion matrix, we can calculate: the recall, precision than we combined the two measures to calculate the F-Measure like that:

$$F - Mesure = \frac{2 * (Précision * Rappel)}{(Précision + Rappel)} \quad (7)$$

3.3 Result

Results of Layer 1 : Before Optimisation with Social Spiders

Phrases score threshold / Similarity		0,60						0,65					
threshold simila-rity	Metric evaluation	REG	Cortex	Humain	Nbr word	Nbr Phrase	Reduced rates	REG	Cortex	Humain	Nbr word	Nbr Phrase	Reduced rates
0,60	ROUGE	0,67	0.71	0.55	245	15	21,72%	0,67	0,68	0,52	224	13	28,43%
	F-Mesure	0,49	0,46	0,32				0.55	0.50	0.47			
0,65	ROUGE	0,65	0,69	0,55	232	14	25,87%	0,72	0,68	0,53	221	12	29,39%
	F-Mesure	0.51	0.49	0.37				0.58	0.54	0.52			
0,70	ROUGE	0,61	0,68	0,51	230	14	26,51%	**0,71**	**0,69**	**0,56**	**208**	**10**	**33,54%**
	F-Mesure	0.57	0.51	0.44				**0.64**	**0.62**	**0.55**			
Phrases score threshold / Similarity		0,70						0,75					
threshold simila-rity	Metric evaluation	REG	Cortex	Humain	Nbr word	Nbr Phrase	Reduced rates	REG	Cortex	Humain	Nbr word	Nbr Phrase	Reduced rates
0,60	ROUGE	**0,73**	**0,74**	**0,55**	**193**	**10**	**38,33%**	0,67	0,67	0,58	123	6	60,70%
	F-Mesure	**0.61**	**0.64**	**0.59**				0.43	0.47	0.45			
0,65	ROUGE	0,67	0,70	0,57	180	8	42,49%	0,68	0,68	0,58	113	5	63,89%
	F-Mesure	0.58	0.59	0.56				0.42	0.45	0.41			
0,70	ROUGE	0,71	0,68	0,54	143	7	54,31%	0,71	0,74	0,64	110	4	64,85%
	F-Mesure	0.57	0.55	0.55				0.39	0.45	0.38			

Fig. 1. Result of Layer 1: Before optimization with Social Spider

- In Yellow: the local optimal candidate summary before optimization, quoted just for illustration, but will not be used for optimization with social spiders.
- In the Green: abstract global optimal candidate before optimization, which will be used for optimization with social spiders,

Results of Layer 3 : After Optimization with the Social Spider and VOTE.

We used two social spiders parameter combined with Number of iterations = 500

	Combine1	Combine 2
Threshold higher discount rate	55% =0,55	50%= 0,50
Threshold lower discount rate	27,5%=0.275	30%=0,30
Number of spiders	3	3
Maximum number of common stake in the same order	5	5

	Metric evaluation	REG	Cortex	Humain	Nbr word	Nbr Phrase	Reduced rates	Execution time
Before Optimization	ROUGE	**0,71**	**0,69**	**0,56**	**208**	**10**	**33,54%**	**819 ms**
	F-Mesure	**0.64**	**0.62**	**0.55**				
Optimizing social spider (Combine 1)	ROUGE	0.72	0.73	0.60	205	9	34,50%	3602 ms
	F-Mesure	0.66	0.67	0.57				
Optimizing social spider (Combine 2)	ROUGE	0.72	0.75	0.61	195	9	37,69%	2762 ms
	F-Mesure	0.68	0.72	0.66				

Fig. 2. Optimization *of 1st summary candidate score threshold=0.*65, threshold of similarity=0.70

	Metric evaluation	REG	Cortex	Humain	Nbr word	Nbr Phrase	Reduced rates	Execution time
Before Optimization	ROUGE	**0,73**	**0,74**	**0,55**	**193**	**10**	**38,33%**	**833 ms**
	F-Mesure	**0.61**	**0.64**	**0.59**				
Optimizing social spider (Combine 1)	ROUGE	0.68	0.66	0.47	187	9	40,25%	3859 ms
	F-Mesure	0.64	0.62	0.55				
Optimizing social spider (Combine 2)	ROUGE	0.75	0.78	0.62	190	9	39,29%	2591 ms
	F-Mesure	0.65	0.66	0.6				

Fig. 3. Optimization *of 2nd summary candidate score threshold=0.*70, threshold of similarity=0.60

We conducted a series of experiments to find and fix the most optimal parameters of social spiders.

3.4 Interpetation

We experimented document "Hurricane" using the coding TFC for the first stage (scoring) and several similarity distances (second stage) to try to detect theUSER sensitive about the best results we summarized validated by the metric RED by comparing the summary reference from REG system COTREX and a human expert who summed us the text "Hurricaine". All tests on data representation parameters were performed to éviterde misjudge our new approach based on a biomimetic approach in this case social spiders.

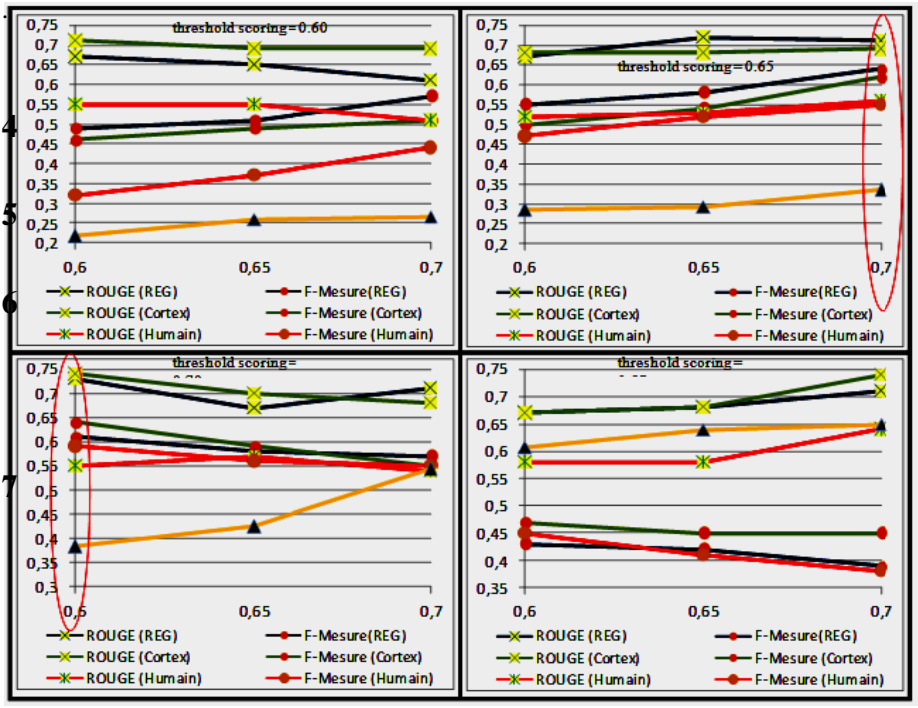


Fig. 4. Summary Evaluation graph before optimization (layer 1)

The first sub-graph (top left corner) indicates incoérence between the two F-evaluation metric measurement and ROUGE incoérence this is resulting from a false assessment of ROUGE summary. This is explained by the weak against the ROUGE summary negligible rate reduction: in fact a summary has low reduction rate will have the lesco-occurrences of N-grams number between him and a set of reference summaries Rref larger than a summary has greatly reduced rates.

The second sub-graph (top right corner) and the third sub-graph (bottom left) sub-graph shows complete coherence between the three evaluation indexes: reduction rate, F-Measure and ROUGE. While the fourth sub.

According to the experimental set of results when we set the target parameter values, it has turn out that:

(a) Increasing number of iterations and the increase in social spiders influences the execution time, the candidate summary quality is not reached by the change of these two parameters
(b) Maximum number of common stake in the same order minimizes the number of abstracts same candidate before the vote and can cover the maximum possible case

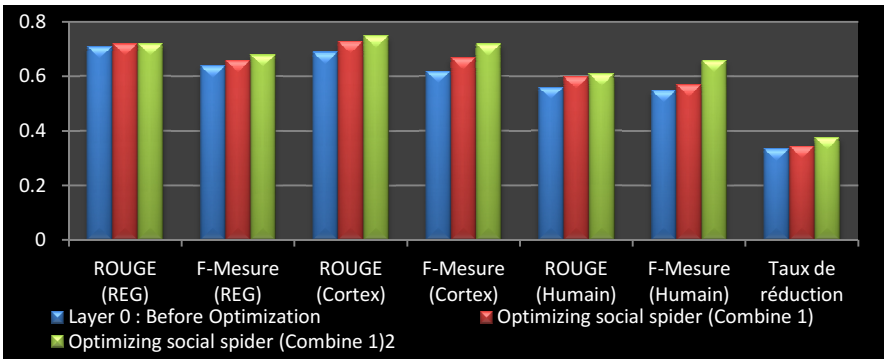


Fig. 5. Optimization of the first summary candidate score threshold=0.65, threshold of similarity= 0.70

The graph below shows explicitly that the second parameter optimization combined with social spiders return results better compared to the first combination, this is explained by the given interval of utility and semantics represented by two thresholds: upper and lower discount rate is reduced, which allows well-directed social spider. While the first combined with a wider interval, that channels less the optimization work.

We note that the execution time optimization combined with the first is greater than the second combines this means that the search field combines 1 is greater than the second combination.

8 Conclusion and Perspective

In this article, we presented new ideas: the first is to have used two techniques of extraction summary after another to improve the rate of reduction without loss of semantics.

The second idea is the use of a biomimetic approach that has the representation of strength graph, social spiders can almost total coverage on a graph using the communication module.

Given the results obtained, our approach based on a biomimetic approach (social spiders) can help solve one of the problems of textual data exploration and visualization will.

Prospects we will try to improve this approach using the WordNet thesaurus, and use a summary based on feelings using the SentiWordNet. We'll also try to explore other biomimetic methods. For nature still has not revealed all the secrets.

Reference

1. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development 2(2), 159–165 (1958)
2. Edmundson, H.P.: Automatic Abstracting, TRW Computer Division. Thompson Ram Wooldridge. Inc., Canoga Park (1963)
3. DeJong, G.: An overview of the FRUMP system. In: Strategies for Natural Language Processing 113 (1982)
4. Fum, D., Guida, G., Tasso, C.: Forward and backward reasoning in automatic abstracting. In: Proceedings of the 9th Conference on Computational Linguistics, vol. 1, pp. 83–88. AcademiaPraha (July 1982)
5. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Information Processing& Management 33(2), 193–207 (1997)
6. Mitra, M., Buckley, C., Singhal, A., Cardie, C.: An Analysis of Statistical and Syntactic Phrases. In: RIAO, vol. 97, pp. 200–214 (June 1997)
7. Teufel, S., Moens, M.: Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: Advances in Automatic Text Summarization, pp. 155–171 (1999)
8. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B.: The TIPSTER SUMMAC text summarization evaluation. In: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, pp. 77–85. Association for ComputationalLinguistics (June 1999)
9. Kim, S.N., Medelyan, O., Kan, M.Y., Baldwin, T.: Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21–26. Association for Computational Linguistics (July 2010)
10. Boudin, F., Morin, E.: Keyphrase Extraction for N-best reranking in multi-sentence compression. In: North American Chapter of the Association for Computational Linguistics (NAACL) (June 2013)
11. Hovy, E., Lin, C.Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006), pp. 604–611 (May 2006)
12. Donaway, R.L., Drummey, K.W., Mather, L.A.: A comparison of rankings produced by summarization evaluation measures. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization, vol. 4, pp. 69–78. Association for Computational Linguistics (April 2000)
13. Cuevas, E., Cienfuegos, M., Zaldívar, D., Pérez-Cisneros, M.: A swarm optimization algorithm inspired in the behavior of the social-spider. Expert Systemswith Applications 40(16), 6374–6384 (2013)
14. Hamou, R.M., Amine, A., Rahmani, M.: A new biomimetic approach based on social spiders for clustering of text. In: Lee, R. (ed.) Software Engineering Research, Management and Appl. 2012. SCI, vol. 430, pp. 17–30. Springer, Heidelberg (2012)
15. Hamou, R.M., Amine, A., Lokbani, A.C.: The Social Spiders in the Clustering of Texts: Towards an Aspect of Visual Classification. International Journal of Artificial Life Research (IJALR) 3(3), 1–14 (2012)

Building Domain Specific Sentiment Lexicons Combining Information from Many Sentiment Lexicons and a Domain Specific Corpus

Hugo Hammer(✉), Anis Yazidi, Aleksander Bai, and Paal Engelstad

Department of Computer Science, Oslo and Akershus University College of Applied Sciences, Oslo, Norway
{hugo.hammer,anis.yazidi,aleksander.bai,paal.engelstad}@hioa.no

Abstract. Most approaches to sentiment analysis requires a sentiment lexicon in order to automatically predict sentiment or opinion in a text. The lexicon is generated by selecting words and assigning scores to the words, and the performance the sentiment analysis depends on the quality of the assigned scores. This paper addresses an aspect of sentiment lexicon generation that has been overlooked so far; namely that the most appropriate score assigned to a word in the lexicon is dependent on the domain. The common practice, on the contrary, is that the same lexicon is used *without adjustments* across different domains ignoring the fact that the scores are normally highly sensitive to the domain. Consequently, the same lexicon might perform well on a single domain while performing poorly on another domain, unless some score adjustment is performed. In this paper, we advocate that a sentiment lexicon needs some further adjustments in order to perform well in a specific domain. In order to cope with these domain specific adjustments, we adopt a stochastic formulation of the sentiment score assignment problem instead of the classical deterministic formulation. Thus, viewing a sentiment score as a stochastic variable permits us to accommodate to the domain specific adjustments. Experimental results demonstrate the feasibility of our approach and its superiority to generic lexicons without domain adjustments.

Keywords: Bayesian decision theory · Cross-domain · Sentiment classification · Sentiment lexicon

1 Introduction

With the increasing amount of unstructured textual information available on the Internet, sentiment analysis and opinion mining have recently gained a groundswell of interest from the research community as well as among practitioners. In general terms, sentiment analysis attempts to automate the classification of text materials as either expressing positive sentiment or negative sentiment. Such classification is particularity interesting for making sense of huge amount of text information and extracting the "word of mouth" from different domains like product reviews, movie reviews, political discussions etc.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 205–216, 2015.
DOI: 10.1007/978-3-319-19578-0_17

There are two main approaches to sentiment classification

- *Sentiment lexicon:* A sentiment lexicon is merely composed of sentiment words and sentiment phrases (idioms) characterized by sentiment polarity, positive or negative, and by sentimental strength. For example, the word 'excellent' has positive polarity and high strength whereas the word 'good' is also positive but has a lower strength. Once a lexicon is built and in place, a range of different approaches can be deployed to classify the sentiment in a text as positive or negative. These approaches range from simply computing the difference between the sum of the scores for the positive lexicon and the sum of the scores for the negative lexicon, and subsequently classifying the sentiment in the text according to the sign of the difference.
- *Supervised learning:* Given a set of documents with known sentiment class, the material can be used to train a model to classify the sentiment class of new documents.

A major challenge in sentiment classification is that the classification method normally is highly sensitive to the domain. A method that performs well in one domain, may not perform well in a different domain. It is worth mentioning that the later problem is common and well studied in the field of Machine Learning, since supervised learning is especially sensitive to the domain, and typically it performs well only in the domain of the annotated documents. The later problem is referred to in the literature as cross-domain classification.

Several methods have been suggested to overcome this challenge in the field of sentiment analysis. However, they are merely inspired by the legacy research on cross-domain classification in the field of machine learning. These methods are often referred to as cross-domain sentiment classification [1]. The premises of these methods is to adjust a supervised classifier to the domain of interest. The approaches consist of either using a small annotated corpus or, alternatively, a large non-annotated corpus from the domain of interest [2,3,4].

In this paper we study another problem, which is very common in practice, but to the best of our knowledge has not been studied in the literature. For many languages several different sentiment lexicons are available, and it is often difficult to know which sentiment lexicon is preferable. Ideally one would like to use the information from all the lexicons, but this is often challenging since the scores of a sentiment word varies between the lexicons and may also be contradictory. In addition there is usually also a large amount of text from the domain of interest, e.g. a large set of product reviews that we want to classify with respect to sentiment. We present a method that builds domain specific sentiment lexicons using information from the sentiment lexicons and the corpus from the domain of interest in an advantageous way. The suggested method is based on Bayesian decision theory.

Before, we proceed to presenting our solution and our experimental results, we shall present a brief review of the related work. Most of the research within cross domain sentiment classification focuses on devising approaches to join information from labelled and/or unlabelled corpuses from different domains and the domain of interest to improve sentiment classification.

Bollegale et al. [5] argue that a major challenge of applying a classifier trained on one domain to another is that features may be quite different in different domains. The authors suggest to develop a sentiment sensitive thesaurus to expand the number of features in both the training and test sets.

Pan et al. [6] consider the case with unlabelled data in the domain of interest and labelled data from an other domain. To bridge the gap between the domains, the authors propose a spectral feature alignment algorithm to align domain-specific words from different domains into unified clusters, with the help of domain independent words as bridges.

Chetviorkin and Loukachevitch [7] propose a statistical features based approach in order to discriminate sentiment words in different domains do develop domain specific sentiment lexicons. The method requires labeled corpuses from both the domain of interest and the other domains.

Contextual sentiment lexicons takes the context of the sentiment words into account. Such lexicons are usually even more sensitive to the domain than ordinary sentiment lexicons are. Gindl et al. [8] suggest a method that identifies unstable contextualizations and refines the contextualized sentiment dictionaries accordingly, eliminating the need for specific training data for each individual domain.

In [9], the authors identified words that exhibit dis-ambiguity based on cross-domain evaluations. In simple terms, if a word gets a positive score in a domain with high confidence and a negative score in another domains, then this terms is considered dis-ambiguous. The next step was to create a domain-independent lexicon by simply excluding the words which are dis-ambiguous across domains. In [10], a taxonomy is used to determine the domain such as movies, politics, sports, then the different lexicons are learned on a domain basis. However, the authors did not discuss adjusting the scores across domains.

2 Joining Information from Sentiment Lexicons and Domain Specific Corpus

Our method consists of two parts. First we join the information from the sentiment lexicons, and second we adjust this information using the domain specific corpus.

2.1 Posterior Expected Sentiment Score

We assume that we have a total of n_L sentiment lexicons consisting of a total of n_W sentiment words occurring in at least one of the sentiment lexicons. We denote the sentiment words $w_1, w_2, \ldots, w_{n_W}$. Let $s_{i,i(j)},\ i = 1, \ldots, n_W,\ j = 1, \ldots, |s_i|$ denote the sentiment score for sentiment word w_i in sentiment lexicon $i(j) \in \{1, 2, \ldots n_L\}$. $|s_i|$ denotes the number of lexicons that word i occurs in, while $i(1), i(2), \ldots, i(|s_i|)$ are references to these lexicons. Naturally $|s_i| \leq n_L,\ i = 1, 2, \ldots, n_W$. We assume that $s_{i,i(j)}, j \in 1, \ldots, |s_i|$ are independent outcomes from $N(\mu_i, \sigma)$ denoting a normal distribution with expectation μ_i and

standard deviation σ. Further we assume that outcomes from different sentiment words are independent. We associate prior distributions to the unknown parameters $\mu_i \sim N(0, \tau)$ and $\sigma^2 \sim \text{InvGamma}(\alpha, \beta)$. From the regression model we can estimate the posterior distributions $P(\mu_i | s_{i,i(1)}, \ldots, s_{i,i(|s_i|)}), i = 1, \ldots, n_W$ which will be used in the next Section.

2.2 Bayesian Decision Theory

In the traditional decision theory we assume that we have a set of stochastic variables $X_1, X_2, \ldots, X_n$ where $X_i \sim f(x|\theta)$ and in the Bayesian framework we assume a prior distribution $\theta \sim p(\theta)$. We want to decide a value for the unknown parameter θ and denote this decision (action) a. In Bayesian decision theory we chose a value a minimizing the posterior expected loss

$$\begin{aligned}\widehat{a} &= \underset{a}{\text{argmin}} \left\{ E_\theta(L(a;\theta)|x_1, x_2, \ldots, x_n) \right\} \\ &= \underset{a}{\text{argmin}} \left\{ \int_\theta L(a;\theta)\, p(\theta|x_1, x_2, \ldots, x_n)\, \mathrm{d}\theta \right\}\end{aligned}$$

where $p(\theta|x_1, x_2, \ldots, x_n)$ is the posterior distribution and $L(a;\theta)$ the loss function that returns the loss of the decision $\theta = a$. The most common loss function is the quadratic loss $L(a;\theta) = (a - \theta)^2$ which results in the action $\widehat{a} = E_\theta(\theta|x_1, x_2, \ldots, x_n)$, the posterior expectation.

2.3 Corpus Loss Function

In this section we join the information from the sentiment lexicons and the domain corpus minimizing the posterior expected loss. Our loss function consists of two parts. The first part is the quadratic loss function based on the sentiment lexicons

$$L_1(a_i; \mu_i) = (a_i - \mu_i)^2$$

The second part of the loss function incorporates information from a corpus from the domain of interest. We assume that the corpus consist of D document and could for example be a large set of product reviews, movie reviews or news articles that we need to classify with respect to sentiment. We assume that the true sentiment classes of these documents are unknown, but still these documents contain valuable sentiment information by the fact that sentiment words in the same document tend to have similar values [11]. For example a positive review typically consists of more positive than negative sentiment words. In traditional sentiment lexicon based classification this valuable information is not used. In the second part of the loss function we incorporate this information setting that the loss increases if a_i differs more from the expected sentiment value of the neighboring sentiment words in the same document

$$L_2(a_i; \mu_1, \mu_2, \ldots, \mu_{n_W}) = \sum_{d=1}^{D} \sum_{k=1}^{N_{id}} \sum_{p=1}^{P_{id}} \frac{1}{\delta(w_{ikd}, \widetilde{w}_{kdp}) + 1} \left[a_i - \psi(w_{ikd}, \widetilde{w}_{dp}) \widetilde{\mu}_{dp} \right]^2$$

where N_{id} is the number of times w_i occurs in document d, w_{ikd} occurrence number k of sentiment word w_i in document d. Further, $\widetilde{w}_{d1}, \ldots, \widetilde{w}_{dP_{id}}$ denote the other occurrences of sentiment words in document d except w_{ikd} and $\widetilde{\mu}_{d1}, \ldots, \widetilde{\mu}_{dP_{id}}$ is the expected sentiment value of these sentiment words according to the model in Section 2.1.

The word 'good' has a positive sentiment while the phrase 'not good' has a negative sentiment. Thus the word 'not' results in a shift in sentiment. Words like 'not', 'never', 'none', 'nobody' are referred to as sentiment shifters [1] and it is natural to change the sentiment of a sentiment word if it is close to a sentiment shifter. The function $\psi(w_{ikd}, \widetilde{w}_{dp})$ includes the sentiment shift in the comparison of w_{ikd} and $\widetilde{w}_{dp}$. If there are no shifters close to either w_{ikd} or $\widetilde{w}_{dp}$ no shift is necessary, and $\psi(w_{ikd}, \widetilde{w}_{dp}) = 1$. If there is a sentiment shifter close to w_{ikd} or close to $\widetilde{w}_{dp}$ the sentiment of one of them is shifting, and thus $\psi(w_{ikd}, \widetilde{w}_{dp})$ is equal to -1. In some rare cases there is more than one sentiment shifter close to w_{ikd} and $\widetilde{w}_{dp}$. We than use the rule that two shifters outweigh each other. Thus, more generally, we use the rule that if in total there is an odd number of sentiment shifters close to w_{ikd} and $\widetilde{w}_{dp}$, then $\psi(w_{ikd}, \widetilde{w}_{dp})$ is equal to -1, or else it is equal to 1.

Finally, the function $\delta(w_{ikd}, \widetilde{w}_{dp})$ returns the number of words between w_{ikd} and $\widetilde{w}_{dp}$. The shorter the distance $\delta(w_{ikd}, \widetilde{w}_{dp})$, the more likely the sentiment values are expected to be similar [12,13]. Thus, we set the loss inversely proportional to the distances $\delta(w_{ikd}, \widetilde{w}_{dp}) + 1$.

The overall loss function is a weighted sum of the two loss functions presented above.

$$L(a_i) = \alpha N_i L_1(a_i) + (1 - \alpha) L_2(a_i), \ \alpha \in [0, 1]$$

where $N_i = \sum_{d=1}^{D} N_{id}$, the number of times w_i occurs in the corpus. With $\alpha = 1$, the loss function only depends on the sentiment lexicons and not on the corpus. The lower the value α, the more the loss function depends on information from corpus ($L_2(a_i)$).

Let $\widehat{a}_i$ denote that value of a_i that minimizes the posterior expected loss

$$\widehat{a}_i = \underset{a_i}{\operatorname{argmin}} \, E\left[L\left(a_i\right)\right]$$

with respect to the posterior distributions of $\mu_i, i = 1, 2, \ldots, n_w$. Straight forward computations gives

$$\widehat{a}_i = \frac{\alpha N_i E_i + (1 - \alpha) \sum_{d=1}^{D} \sum_{k=1}^{N_{id}} \sum_{p=1}^{P_{id}} \frac{\psi(w_{ikd}, \widetilde{w}_{dp})}{\delta(w_{ikd}, \widetilde{w}_{dp}) + 1} \widetilde{E}_{dp}}{\alpha N_i + (1 - \alpha) \sum_{d=1}^{D} \sum_{k=1}^{N_{id}} \sum_{p=1}^{P_{id}} \frac{1}{\delta(w_{ikd}, \widetilde{w}_{dp}) + 1}}$$

where E_i denote the posterior expectation $E(\mu_i | s_{i,i(1)}, \ldots, s_{i,i(|s_i|)})$ and similarly $\widetilde{E}_{dp}$ is the posterior expectation of $\widetilde{\mu}_{dp}$. In accordance with Section 2.2, with $\alpha = 1$ the sentiment value $\widehat{a}_i$ becomes equal to the posterior expectation, E_i.

3 Preexisting Sentiment Lexicons

For the method in Section 2 we use three different sentiment lexicons developed for the Norwegian language.

Translation. The first sentiment lexicon was generated by translating the well-known English sentiment lexicon AFINN [14] to Norwegian using machine translation (Google translate) and doing further manual improvements. We denote this lexicon AFINN in the rest of the paper.

Synonym Antonym Word Graph. To create the second sentiment lexicon we first built a large undirected graph of synonym and antonym relations between words from three Norwegian thesauruses. The words were nodes in the graph and synonym and antonym relations were edges. The full graph consists of a total of 6036 nodes (words), where 109 of the nodes represent the seed words (51 positive and 57 negative), and there are 16475 edges (synonyms and antonyms) in the graph. The seed words were manually selected, picking words that are used frequently in the Norwegian language and that span different dimensions of both positive sentiment ('happy', 'clever', 'intelligent', 'love' etc.) and negative sentiment ('lazy', 'aggressive', 'hopeless', 'chaotic' etc.). The sentiment lexicon was generated using the Label Propagation algorithm [15], which is the most common algorithm for this task. The initial phase of the Label Propagation algorithm consists of giving each positive and negative seed a word score 1 and -1, respectively. All other nodes in the graph are given score 0. The algorithm propagates through each non-seed words updating the score using a weighted average of the scores of all neighbouring nodes (connected with an edge). When computing the weighted average, synonym and antonym edges are given weights 1 and -1, respectively. The algorithm is iterated until changes in scores are below some threshold for all nodes. The resulting score for each node becomes our derived sentiment lexicon. For more details, we refer the reader to our previous work [16]. We denote this sentiment lexicon LABEL in the rest of the paper.

From Corpus. The third sentiment lexicon was constructed using the corpus based approach [17] on a large Norwegian corpus consisting of about one billion words. We started with 14 seed words, seven with positive and seven with negative sentiment and computed the Pointwise mutual information (PMI) between the seed words and the 5000 most frequent words in the corpus and 8340 adjectives not being part of the 5000 most frequent words. The computed PMI scores lay the foundation for the sentiment lexicon. For more details, see [18]. We denote this lexicon PMI in the rest of the paper.

Based on the sentiment lexicons described above, we generated three sentiment lexicons using the method in Section 2 with $\alpha = 0, \alpha = 0.5$ and $\alpha = 1$. In the rest of the paper we denote these sentiment lexicons W0, W0.5 and W1, respectively. We adjusted the sentiment lexicons towards the domain of product reviews using the text from 15118 product reviews from the Norwegian online shopping sites `www.komplett.no`, `mpx.no`. For the sentiment shifter function $\psi(w_{ikd}, \widetilde{w}_{dp})$ in the loss function L_2 in Section 2.3 recall that the sentiment of

a sentiment word is shifted if a sentiment shifter is close to the sentiment word. In the computations in this paper we decided to shift sentiment if the sentiment shifter was one or two words in front of the sentiment word. We only used the sentiment shifter 'not' ('ikke'), but also considered other sentiment shifters, such as 'never' ('aldri'), and other distances between the sentiment word and the shifter. However, the selected approach presented in this paper seems to be the best for such lexicon approaches in Norwegian [19].

4 Evaluating Classification Performance

For each of the product reviews from `www.komplett.no` and `mpx.no` a rating from 1 to 5 is known and is used to evaluate the classification performance of each of the sentiment lexicon described above.

For each lexicon, we computed the sentiment score of a review by simply adding the score of each sentiment word in a sentiment lexicon together, which is the most common way to do it [20]. Similar as for the sentiment shifter function $\psi(w_{ikd}, \widetilde{w}_{dp})$ in L_2 we shifted the sentiment of a sentiment word if the sentiment shifter 'not' ('ikke') was one or two words in front of the sentiment word. Finally the sum is divided by the number of words in the review, giving us the final sentiment score for the review.

Classification Method. We divided the reviews in two equal parts, one half being training data and the other half used for testing. We used the training data to estimate the average sentiment score of all reviews related to the different ratings. The computed scores could look like Table 1. We classified a review from

Table 1. Average computed sentiment score for reviews with different ratings

Rating	1	2	3	4	5
Average sentiment score	−0.23	−0.06	0.04	0.13	0.24

the test set using the sentiment lexicon to compute a sentiment score for the test review and classify to the closest average sentiment score from the training set. E.g. if the computed sentiment score for the test review was −0.05 and estimated averages were as given in Table 1, the review was classified to rating 2. In some rare cases the estimated average sentiment score was not monotonically increasing with the rating. Table 2 shows an example where the average for rating 3, is higher than for the rating 4. For such cases, the average of the two

Table 2. Example were sentiment score were not monotonically increasing with rating

Rating	1	2	3	4	5
Average sentiment score	−0.23	−0.06	0.18	0.10	0.24

sentiment scores were computed, $(0.10 + 0.18)/2 = 0.14$, and classified to 3 or 4 if the computed sentiment score of the test review was below or above 0.14, respectively.

Classification Performance. We evaluated the classification performance using average difference in absolute value between the true and predicted rating for each review in the test set

$$\text{Average abs. error} = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i|$$

where n is the number off reviews in the test set and p_i and r_i is the predicted and true rating of review i in the test set. Naturally, a small average absolute error would mean that the sentiment lexicon performs well.

Note that the focus in this paper is not to do a best possible classification performance based on the training material. If that was our goal, other more advanced and sophisticated techniques would be used, such as machine learning based techniques. Our goal is rather to evaluate and compare the performance of sentiment lexicons, and the framework described above is chosen with respect to that.

5 Results

This section presents the results of classification performance on product reviews for the different sentiment lexicons. The results are shown in Table 3. Training

Table 3. Classification performance for sentiment lexicons on `komplett.no` and `mpx.no` product reviews. The columns from left to right show the sentiment lexicon names, the number of words in the sentiment lexicons, mean absolute error with standard deviation and 95% confidence intervals for mean absolute error.

	N	Mean (Stdev)	95% conf.int.
AFINN	2260	1.17 (1.11)	(1.14, 1.19)
W0.5	14987	1.24 (1.17)	(1.22, 1.27)
W1	14987	1.29 (1.17)	(1.26, 1.31)
LABEL	6036	1.38 (1.27)	(1.36, 1.41)
W0	14987	1.52 (1.37)	(1.49, 1.55)
PMI	13340	1.53 (1.34)	(1.50, 1.56)

and test sets were created by randomly adding an equal amount of reviews to both sets. All sentiment lexicons were trained and tested on the same training and test sets, making comparisons easier. This procedure was also repeated several times, and every time the results were in practice identical to the results in Tables 3, documenting that the results are independent of which reviews that were added to the training and test sets.

Recall that we constructed W0, W0.5 and W1 based on the lists AFINN, LABEL and PMI which we call the source lexicons in the rest of this paper. We see that the source lexicons varies quite much in performance, ranging from 1.17 to 1.53, with the AFINN lexicon being the best. This indicates that translation of sentiment lexicons from one language to another can be an efficient way to construct viable sentiment lexicons (at least when the languages are related, such as the two Germanic, Indo-European languages, English and Norwegian.) Both of the sentiment lexicons that solely rely on corpus (PMI and W0) perform poorer than the other sentiment lexicons. Even though the performance of the source lexicons varies quite much, the performance of W0.5 and W1 is very good and almost as well as the best of the source lexicons (AFINN) and much better than the two other source lexicons (LABEL and PMI). Interestingly W0.5 performs significantly better than both W1 (paired T-test p-value = 0.022) and W0 (p-value = $2.3 \cdot 10^{-7}$) showing that the best sentiment lexicon is the one that is constructed by combining the information from both the source sentiment lexicons and the product review corpus.

Tables 4 and 5 show sentiment words that have the largest difference in sentiment score between the two sentiment lexicons W0 and W1 and that occur at least 50 times in the product review corpus. These were the sentiment words that were adjusted the most when the information from the product review corpus were included. Similar to other corpus based methods, noise is introduced, and we observe examples of this noise in the tables. E.g. we see that words like 'fabulous' and 'awesome' have been changes from a positive score to negative/neutral and that words like 'jerk', 'dirty' and 'damn' have been changed from a negative score to positive/neutral. On the other hand, we also see several words that

Table 4. Sentiment words where the sentiment scores are decreased the most when the information from the corpus is included. Columns from left to right: Sentiment words in Norwegian, in English, sentiment scores in the sentiment lexicons W0 and W1 and the difference between these sentiment scores.

Norwegian	English	Lexicon W1	Lexicon W0	Difference
skada	damaged	−0.35	1.78	−2.13
gult	yellow	0.19	2.26	−2.07
forklarer	explains	−0.33	1.68	−2.01
rikelig	plenty	−0.45	1.53	−1.98
fabelaktige	fabulous	−0.33	1.61	−1.93
knotete	tricky	0.14	2.07	−1.93
fantastisk	awesome	0.20	2.11	−1.91
dårligt	bad	−0.09	1.82	−1.91
søt	sweet	−0.43	1.31	−1.74
forholdet	relationship	−0.07	1.66	−1.73
jublet	cheered	−0.47	1.26	−1.73
finale	finale	−0.07	1.65	−1.73
anvendelig	applicable	0.29	2.00	−1.71
kontakter	contacts	0.01	1.66	−1.65

Table 5. Sentiment words where the sentiment scores are increased the most when the information from the corpus is included. Columns from left to right: Sentiment words in Norwegian, in English, sentiment scores in the sentiment lexicons W0 and W1 and the difference between these sentiment scores.

Norwegian	English	Lexicon W1	Lexicon W0	Difference
vinne	win	1.19	−1.20	2.39
nedsatt	reduced	0.37	−1.84	2.22
angitt	specified	0.37	−1.81	2.19
vunnet	won	0.79	−1.37	2.15
sjokkerende	shocking	0.85	−1.29	2.14
reklamerte	advertised	0.18	−1.91	2.09
dust	jerk	0.74	−1.29	2.03
skittent	dirty	0.23	−1.75	1.99
akseptabel	acceptable	0.34	−1.48	1.81
jævlig	damn	0.06	−1.75	1.81
misvisende	misleading	0.22	−1.55	1.76
sensitiv	sensitive	0.05	−1.68	1.73
jenter	girls	0.27	−1.43	1.70
uregelmessig	irregular	0.03	−1.67	1.70
alminnelige	general	0.37	−1.30	1.68

seem to have been changed to a more reasonable score. E.g. we see that words like 'damaged', 'tricky', 'bad', and 'contacts' are changed from a positive score to a negative/neutral value. There are also examples of words that seem to be changed in a reasonable way with respect to the domain of product reviews. E.g. the word 'reduced' is in many contexts a word with negative sentiment, but with respect to product reviews the word is mostly used to state that prices are reduced, which is a positive statement. In Table 5, we see that the word is changed from a negative to a positive sentiment score when the corpus is included.

6 Conclusions

In this paper, we have developed a method to construct domain specific sentiment lexicons by combining the information from many pre-existing sentiment lexicons with an unanotated corpus from the domain of interest. Trying to combine this sources of information has not been investigated in the literature earlier.

In order to cope with these domain specific adjustments, we adopt a stochastic formulation of the sentiment score assignment problem instead of the classical deterministic formulation. Our approach is based on minimizing the expected loss of a loss function that punishes deviations from the scores of the source sentiment lexicons and inhomogeneity in sentiment scores for the same review.

Our results show that a lexicon that combines information from both the source sentiment lexicons and the domain specific corpus performs better than a lexicon that only rely on information from the source lexicons. This lexicon shows an impressive performance that is almost as good as the best of the source lexicons.

References

1. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Toronto (2012)
2. Aue, A., Gamon, M.: Customizing sentiment classifiers to new domains: A case study. In: Proceedings of Recent Advances in Natural Language Processing (RANLP) (2005)
3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics (ACL) (2007)
4. Tan, S., Wu, G., Tang, H., Cheng, X.: A novel scheme for domain-transfer problem in the context of sentiment analysis. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 979–982. ACM, New York (2007), http://doi.acm.org/10.1145/1321440.1321590, doi:10.1145/1321440.1321590
5. Bollegala, D., Weir, D., Carroll, J.: Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus. IEEE Transactions on Knowledge and Data Engineering 25(8), 1719–1731 (2013)
6. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain Sentiment Classification via Spectral Feature Alignment. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 751–760. ACM, New York (2010)
7. Chetviorkin, I., Loukachevitch, N.V.: Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In: COLING, pp. 593–610 (2012)
8. Gindl, S., Weichselbraun, A., Scharl, A.: Cross-Domain Contextualization of Sentiment Lexicons. In: Coelho, H., Studer, R., Wooldridge, M. (eds.) ECAI. Frontiers in Artificial Intelligence and Applications, vol. 215, pp. 771–776. IOS Press (2010)
9. Weichselbraun, A., Gindl, S., Scharl, A.: Extracting and grounding context-aware sentiment lexicons. IEEE Intelligent Systems 28(2), 39–46 (2013)
10. Owsley, S., Sood, S., Hammond, K.J.: Domain specific affective classification of documents. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 181–183 (2006)
11. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424 (2002)
12. Ding, X., Liu, B., Yu, P.S.: A Holistic Lexicon-based Approach to Opinion Mining. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM 2008, pp. 231–240. ACM, New York (2008)
13. Chetviorkin, I., Loukachevitch, N.: Two-Step Model for Sentiment Lexicon Extraction from Twitter Streams. In: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 90–96. Association for Computational Linguistics (2014)
14. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. CoRR abs/1103.2903 (2011)
15. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University (2002)
16. Hammer, H., Bai, A., Yazidi, A., Engelstad, P.: Building sentiment lexicons applying graph theory on information from three Norwegian thesauruses. In: Norweian Informatics Conference (2014)

17. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) 21(4), 315–346 (2003)
18. Bai, A., Hammer, H.L., Yazidi, A., Engelstad, P.: Constructing sentiment lexicons in Norwegian from a large text corpus. In: The 17th IEEE International conference on Computational science and Engineering (CSE), pp. 231–237 (2014)
19. Hammer, H.L., Solberg, P.E.: vrelid, L.O.: Sentiment classification of online political discussions: A comparison of a word-based and dependency-based method. In: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, pp. 90–96 (2014)
20. Bing, L.: Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. Springer (2011)

Improved Cuckoo Search Algorithm for Document Clustering

Saida Ishak Boushaki[1(✉)], Nadjet Kamel[2], and Omar Bendjeghaba[3]

[1] LRIA (USTHB) and University of Boumerdes, Boumerdes, Algeria
saida_2005_compte@yahoo.fr
[2] LRIA (USTHB) and University of Ferhat Abas Setif, Sétif, Algeria
nkamel@usthb.dz
[3] LREEI (UMBB) and University of Boumerdes, Boumerdes, Algeria
benomar75@yahoo.fr

Abstract. Efficient document clustering plays an important role in organizing and browsing the information in the World Wide Web. K-means is the most popular clustering algorithms, due to its simplicity and efficiency. However, it may be trapped in local minimum which leads to poor results. Recently, cuckoo search based clustering has proved to reach interesting results. By against, the number of iterations can increase dramatically due to its slowness convergence. In this paper, we propose an improved cuckoo search clustering algorithm in order to overcome the weakness of the conventional cuckoo search clustering. In this algorithm, the global search procedure is enhanced by a local search method. The experiments tests on four text document datasets and one standard dataset extracted from well known collections show the effectiveness and the robustness of the proposed algorithm to improve significantly the clustering quality in term of fitness function, f-measure and purity.

Keywords: Document clustering · Vector space model · Cuckoo search · Cosine similarity · F-measure · Purity · Metaheuristic · Optimization

1 Introduction

The high advance of the internet has led to exponential growth of the amount of information available in the World Wide Web (WWW). Consequently, exploring the data and finding the relevant information on the web became hard tasks. Over the past decades, many approaches have been developed in order to manage and organize efficiently this large set of documents. For this purpose, clustering is the well known method used by the scientific community dealing by the datamining. It is unsupervised technique [1] [2], that extract hidden structural characteristics in the data and gathering the highly similar objects in the same group, whereas segregates dissimilar objects in different ones. Due to the importance task of the clustering technique, it has been applied in variety engineering and field like image segmentation, pattern recognition and gene-expression. The algorithms of clustering are divided into different categories: hierarchical clustering algorithms, nominal data

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 217–228, 2015.
DOI: 10.1007/978-3-319-19578-0_18

clustering, density based clustering, cohonen networks and partitioning relocation clustering. The last category of clustering contains algorithms with linear time complexity. This makes the partitional algorithms more suitable for web clustering. One of the most famous partitional algorithms is the K-means [3] due to its simplicity and efficiency. However, this algorithm may give a poor results this is due to its random initialization and local exploration, which leads to a local minimum. Actually, nature inspired algorithms cope the shortcoming of a local solution by a global one [4]. One of the most recent metaheuristic algorithms is cuckoo search (CS) optimization [5] [6]. It is based on the interesting breeding behaviour such as brood parasitism of certain species of cuckoos and typical characteristics of Lévy flights. The results of experiment comparison show that the cuckoo search algorithm outperform the most famous metaheuristics [7] [8] [9].

In order to improve the clustering result, and inspired from the hybrid algorithm proposed in [10], we propose in this paper a new algorithm for document clustering, based on CS. In this algorithm, CS is enhanced by additional functions. Which make it superior to conventional CS in term of fitness, convergence speed and external quality.

The remaining of this paper is organized as follows: in section 2, we present most recent metaheuristics algorithms proposed for web document clustering. In section 3, the formal definitions of document clustering are presented. In section 4, we present the fundamental steps of a cuckoo search algorithm for the clustering problem. The improved cuckoo search adapted for document clustering is presented in section 5. Numerical experimentation and results are provided in Section 6. Finally, the conclusion and future work are drawn in Section 7.

2 Related Works

Document clustering based on nature inspired algorithms is an active research field. In 2013, Kamel et al. [11] overcome the weakness of K-means in the initial seed by a hybrid algorithm based on K-means, PSO and Sampling algorithms for document clustering. Leticia Cagnina et al. [12] have presented an improved version of the discrete particle swarm optimization (PSO) algorithm. This version includes a different representation of particles, a more efficient evaluation of the function to be optimized and some modifications in the mutation operator. In 2014, Wei Song et al. proposed a fuzzy control genetic algorithm (GA) in conjunction with a novel hybrid semantic similarity measure for document clustering. It outperforms the conventional GA [13]. In 2013, A novel document clustering algorithm based on ant colony optimization algorithm was proposed by Kayvan Azaryuon and Babak Fakhar. It improves the standard ants clustering algorithm efficiency by making ant movements purposeful, and on the other hand, by changing the rules of ant movement [14]. S. Siamala Devi et al. have used the hybrid K-means with harmony search (HS) to do the comparison between the concept called coverage factor and the concept factorization method for document clustering problem [15]. The experimental results show that factorization produces better results. Recently, the experimental results of

[7] shown that the cuckoo search (CS) clustering achieves best results compared to the well known and recent algorithms: K-means, particle swarm optimization, gravitational search algorithm, the big bang–big crunch algorithm and the black hole algorithm. More recently, in our previous work, we have proposed a new hybrid algorithm for document clustering based on CS and K-means [10]. This new hybrid algorithm outperforms the CS and K-means in term of fitness and external quality.

3 Formal Definitions

Let S be a set of n objects O_1, O_2,..., O_n, each object is defined in multi dimensional space. Clustering S into k clusters means dividing it into k groups or clusters C_1, C_2,..., C_k, such that:

$$\begin{cases} C_i \neq \{\} & \text{for i} = 1,..,\text{k} \\ C_i \cap C_j = \{\} & \text{for i} = 1,..,\text{k, j} = 1,..,\text{k and i} \neq \text{j} \\ C_1 \cup C_2 \cup .. \cup C_K = S \end{cases} \quad (1)$$

In addition, the objects in the same cluster are similar and the objects in different clusters are dissimilar. This property is proportional to the quality of the clustering.

In our case, data are documents. They are represented by using the vector space model (VSM) [10] [16].

The cosine distance is the most used and the best one for document clustering [17]. Given two documents d_i and d_j represented by two vectors v_i and v_j, respectively, the cosine distance is given by the following formula:

$$\cos(d_i, d_j) = \frac{v_i^t v_j}{|v_i||v_j|} \quad (2)$$

Where $|v_i|$ is the norm of the vector v_i

To evaluate the quality of clustering results, we have used two external quality indexes: the famous F-measure and Purity [2] [10].

4 Cuckoo Search Clustering Algorithm

For solving the clustering problem, the standard cuckoo search algorithm is adapted to reach the centroids of the clusters that optimize predefined fitness function. We have used the fitness function presented in [18]. The goal of this function is to find the

solution that maximizes the similarity between each document and the centroid of the cluster that is assigned to. This objective function is given by the following formula:

$$\text{Fitness} \quad \text{Maximize} \sum_{i=1}^{k} \sum_{d_l \in C_i} \cos(d_l, c_i) \tag{3}$$

Where: k is the number of clusters and $\cos(d_l, c_i)$ is the cosine distance between the document d_l and the nearest centroid c_i of the cluster C_i.

The cuckoo search clustering algorithm (CSCA) is given by the following steps [7] [10]:

1. Generate randomly the initial population of *nb_nest* host nests;
2. Calculate the fitness of these solutions and find the best solution;
3. **While (t < *Max_Iter*) or (stop criterion);**
   (a) Generate *nb_nest* new solutions with the cuckoo search;
   (b) Calculate the fitness of the new solutions;
   (c) Compare the new solutions with the old solutions, if the new solution is better than the old one, replace the old solution by the new one ;
   (d) Generate a fraction (p_a) of new solutions to replace the worse nests;
   (e) Compare these solutions with the old solutions. If the new solution is better than the old solution, replace the old solution by the new one;
   (f) Find the best solution;
4. **End while;**
5. Print the best nest and fitness;

5 Improved Cuckoo Search Clustering Algorithm

Cuckoo search clustering algorithm can achieve the best global solution compared to most other metaheuristics. Usually, this global solution is obtained after huge number of iterations due to the slow convergence of the algorithm. It is obvious that in CSCA, the research area is explored using the standard cuckoo function [19]. In the present work, we propose to perform after each new solution generated by the standard cuckoo function an auxiliary local research in the research area in order to improve the solution. If this local search finds a solution that is better than the existing one, then it will be replaced by the new reached one.

For each current solution (host nest), the local search procedure exploits this one by calculating the gravity center of each cluster using the equation (4). Thus, the solutions are replaced only if their new fitness is better. The pseudo code of the new improved cuckoo search clustering algorithm (ICSCA) is presented in Fig. 1.

$$gc_i = \frac{1}{n_i} \sum_{\forall d_l \in C_i} d_l \tag{4}$$

Where gc_i is the gravity center of the cluster C_i, d_l denotes the document that belong to the cluster C_i and n_i is the number of documents in cluster C_i.

```
Begin
1. Set the initial parameters:
 - p_a (the probability of worse nests)
 - nb_nest (the number of host nest is the population size)
 - k (number of clusters)
 - Max_Iter (the maximum number of iterations)
2. Generate randomly the initial population of nb_nest host
   nests;
3. For each solution change the empty clusters
4. Calculate the fitness of each solution using the
   equation(3)and find the best nest;
5. While (t < Max_Iter) or (stop criterion)
 5.1 Generate nb_nest new solutions using the standard
     cuckoo search function;
 5.2 For each new solution change the empty clusters;
 5.3 Calculate the fitness of each new solution using the
     equation (3);
 5.4 For each solution compare the new solutions with the
     old solutions, if the new solution is better than the
     old one, replace the old solution by the new one ;
 5.5 Generate nb_nest new solutions by calculating the
     gravity center of each cluster using equation (4)
 5.6 For each new solution change the empty clusters;
 5.7 Calculate the fitness of each new solution using the
     equation (3);
 5.8 For each solution compare the new solutions with the
     old solutions, if the new solution is better than the
     old one, replace the old solution by the new one;
 5.9 Generate a fraction (p_a) of new solutions to replace
     the worse nests;
 5.10 For each new solution change the empty clusters;
 5.11 Calculate the fitness of each new generated solution
      using the equation (3);
 5.12 Compare the new solutions with the old solutions, if
      the new solution is better than the old one, replace
      the old solution by the new one ;
 5.13 Find the best solution;
End while;
6. Print the best nest and fitness;
End
```

Fig. 1. ICSCA procedure

To illustrate this idea, we give an example. In Fig. 2, we have three clusters and we can see that the objects are more similar to their gravity center than to the centroid generated by the standard cuckoo function.

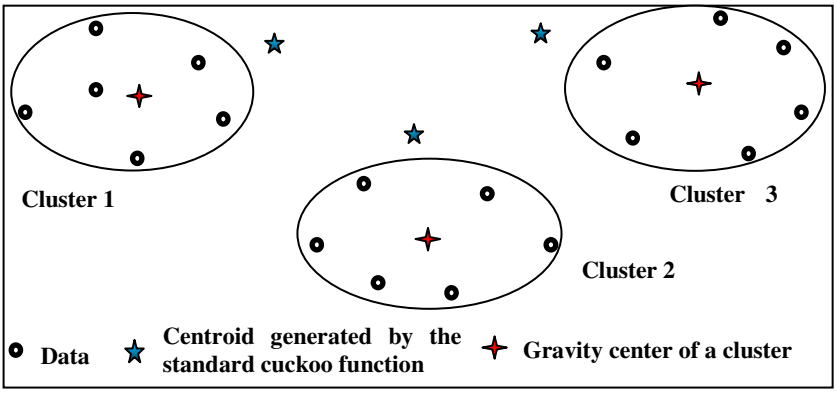


Fig. 2. Example of local search

To illustrate this idea, we give an example. In Fig. 2, we have three clusters and we can see that the objects are more similar to their gravity center than to the centroid generated by the standard cuckoo function.

We should notice that another primary function must be performed after each new generated solution. The main goal of this function is to ensure that there is no empty cluster. The simple way for doing this is to replace the empty cluster by a random one.

6 Experiments and Results

In order to test the efficiency of each auxiliary function added to the standard cuckoo search clustering algorithm, we compare between three algorithms: standard cuckoo search clustering algorithm (CSCA), standard cuckoo search algorithm augmented by the change empty cluster function (CSDC+CEC) and the improved cuckoo search document clustering (ICSCA) enhanced by the local search procedure and the change empty cluster function.

6.1 Datasets

Two kinds of datasets are used in the whole of experiments: four text document datasets and one standard dataset. The text document datasets are extracted from two well known collections: Classic3 [20] and Text REtrieval Conference (TREC) collections [21]. The description detail of text document datasets is given in Table 1,

where the standard dataset is obtained from the famous UCI Machine Learning Repository. The description detail of standard dataset is given in Table 2.

Table 1. Summary of text document datasets

datasets	Number of documents	Number of terms	Classes description	Number of groups
Classic300	300	5471	100, 100, 100	3
Classic400	400	6205	100, 100, 200	3
Tr23	204	5833	6, 11, 15, 36, 45, 91	6
Tr12	313	5805	9, 29, 29, 30, 34, 35, 54, 93	8

Table 2. Description of standard dataset

datasets	Number of instances	Number of attributes	Classes description	Number of groups
Iris	150	4	50, 50, 50	3

6.2 Related Parameters

For the purpose of comparison, the number of iterations is fixed to 100 iterations for the text datasets and only 20 iterations for the Iris standard dataset. We note that for all runs, the probability of worse nests was set to 0.25, while the population size was set to 10. The cosine distance is used as similarity measure for all experiment tests.

6.3 Results and Comparisons

The three algorithms: (CSCA), (CSDC+CEC) and (ICSCA) are compared for the different datasets in term of best fitness value and two external validity indexes (F-measure and purity). In Table 3 we present the best fitness value of the three algorithms for each datasets.

As we can see from this table, the ICSCA can reach the best results in comparison with the CSCA and CSCA+CEC. In addition, the CSCA+CEC is better than CSCA and the gap between them is proportional to the number of clusters. As the number of cluster increases, more than the gap increases.

Table 3. Best fitness value

Datasets	CSCA	CSCA+CEC	ICSCA
Classic300	28.0540	28.3145	56.3282
Classic400	36.7592	36.8108	70.5629
Tr23	29.1706	59.4068	88.5799
Tr12	35.7372	41.5007	93.4783
Iris	149.6669	149.7503	149.8383

For each datasets, the convergence behaviors in term of fitness function obtained by the different algorithms are illustrated in Fig. 3, Fig. 4, Fig. 5, Fig. 6 and Fig. 7.

From these figures, it is clear that the ICSCA can reach the best results in a few iterations number for all datasets. Also, we should notice that the gap between graph variation obtained by the CSCA, and CSCA+CEC algorithms is proportional to the number of cluster. In fact, they are close to each other for Classic300 dataset and Classic400. This is due to the small probability of empty cluster. However, for the Tr12 dataset the gap is more significant due to the big number of clusters.

From Fig. 7, it is obvious that the proposed algorithm speed up the convergence behavior of fitness function. Thus, the cosine distance is accurate for the clustering of the standard dataset.

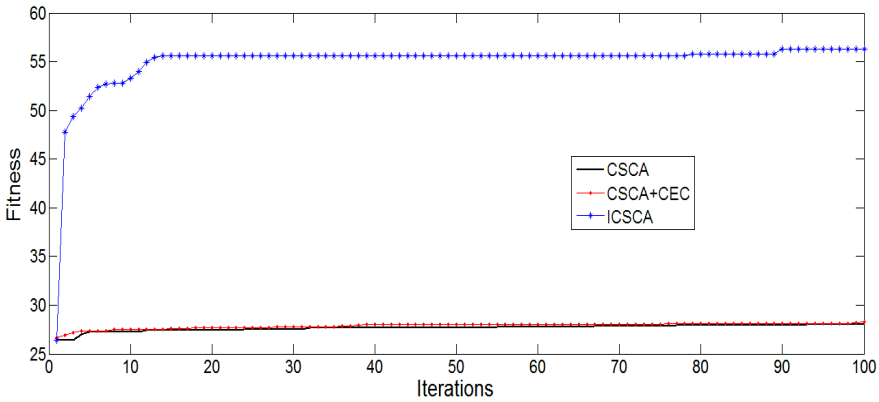


Fig. 3. Graph variation of fitness function of Classic300

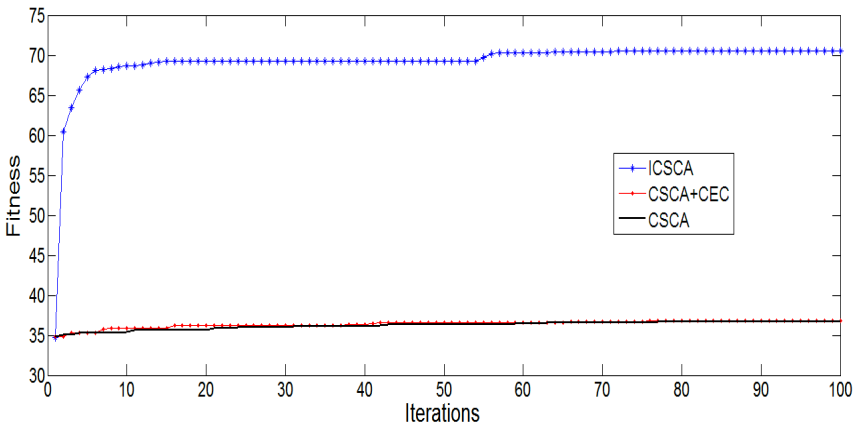


Fig. 4. Graph variation of fitness function of Classic400

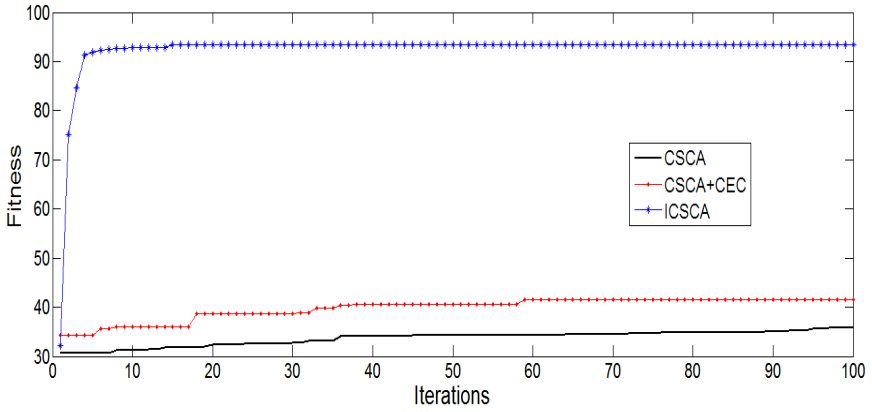


Fig. 5. Graph variation of fitness function of Tr12

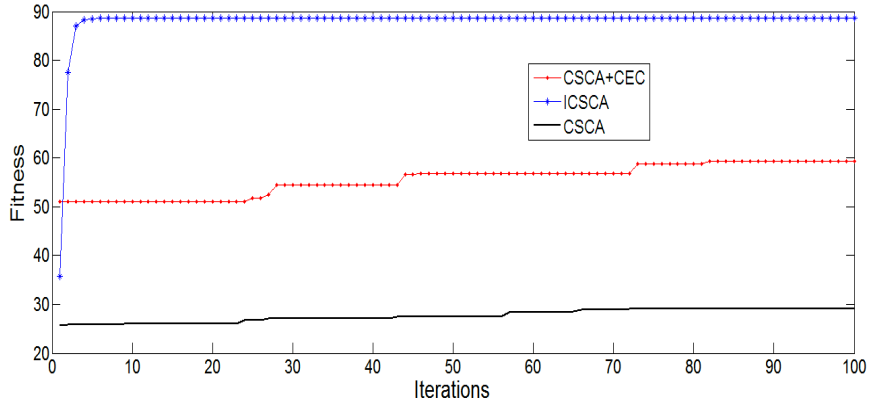


Fig. 6. Graph variation of fitness function of Tr23

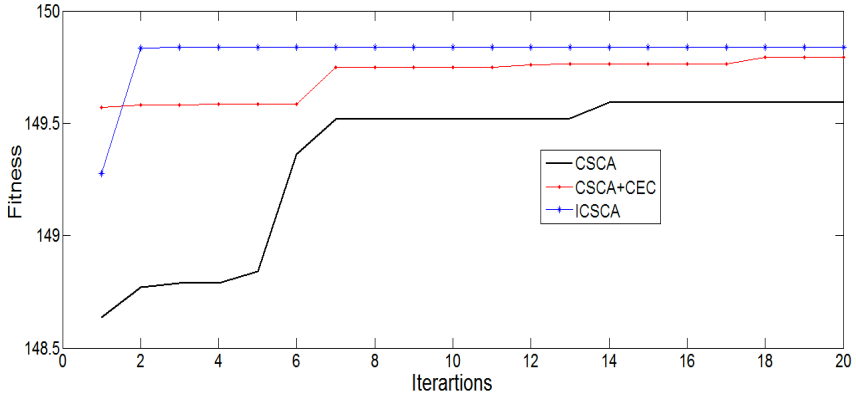


Fig. 7. Graph variation of fitness function of Iris

The recorded F-measure and purity by the different algorithms for each dataset is given in Table 4. and Table 5. From these tables, it is clear that the proposed algorithm can improve significantly the quality of the clustering results.

Table 4. F-measure comparison of CSCA, CSCA+CEC and ICSCA on the differents datasets

Datasets	CSCA	CSCA+CEC	ICSCA
Clasic300	0.3800	0.4160	0.7728
Classic400	0.4109	0.4308	0.6878
Tr23	0.3997	0.4636	0.5476
Tr12	0.2851	0.4187	0.6017
Iris	0.7778	0.9131	0.9666

Table 5. Purity comparison of CSCA, CSCA+CEC and ICSCA on the differents datasets

Datasets	CSDC	CSDC+CCN	ICSDC
Classic300	0.3854	0.3895	0.7863
Classic400	0.3933	0.4071	0.6468
Tr23	0.3138	0.4672	0.4710
Tr12	0.3923	0.4865	0.6532
Iris	0.6667	0.9158	0.9697

The calculated percents that ICSCA improve upon the CSCA+CEC, in terms of fitness function (CPF), f-measure (CPFM) and purity (CPP) is presented in Table 6. It can be stated from this table that the proposed ICSCA is more effective than the CSCA+CEC.

Table 6. Percents improvements of ICSCA improve upon the CSCA+CEC

Datasets	CPF(%)	CPFM(%)	CPP (%)
Classic300	0.4973	0.4616	0.5046
Classic400	0.4783	0.3736	0.3705
Tr23	0.3293	0.1533	0.0080
Tr12	0.2552	0.3041	0.2552
Iris	0,0022	0,0553	0,0555

7 Conclusion

The paper presents an improved cuckoo search clustering algorithm (ICSCA). The novelty of the proposed algorithm is to enhance the conventional cuckoo search clustering by a local search procedure. The experiment results show that the proposed ICSCA is more robust than the CSCA+CEC and CSCA, in term of fitness value, f-measure and purity, when applied on four well known text document dataset and Iris standard dataset. Furthermore, the percent improvement of ICSCA upon the CSCA+CEC is significant.

The proposed ICSCA can also speed up significantly the convergence behavior when applied on Iris standard dataset. Therefore, the cosine distance is accurate for the clustering of the standard dataset. Finally, as future work, we plan to extend the proposed approach for the incremental document clustering.

References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys (CSUR) 31(3), 264–323 (1999)
2. Patel, D., Zaveri, M.: A Review on Web Pages Clustering Techniques. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) NeCoM/WeST/WiMoN 2011. CCI, vol. 197, pp. 700–710. Springer, Heidelberg (2011), doi: 10.1007/978-3-642-22543-7_72.
3. Huang, X., Su, W.: An Improved K-means Clustering Algorithm. Journal of Networks 9(1), 161–167 (2014), doi:10.4304/jnw.9.01.161-167.
4. Hruschka, E.R., Campello, R.J.G.B., Freitas, A., et al.: A Survey of Evolutionary Algorithms for Clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 39(2), 133–155 (2009), doi:10.1109/TSMCC.2008.2007252
5. Yang, X.-S., Deb, S.: Cuckoo Search via Levy Flights. In: World Congress on Proceedings of World Congress on Nature & Biologically Inspired Computing, NaBIC 2009, December 9-11, pp. 210–214. IEEE Publications, Coimbatore (2009), doi:10.1109/NABIC.2009.5393690
6. Yang, X.-S., Deb, S.: Engineering Optimisation by Cuckoo Search. International Journal of Mathematical Modelling and Numerical Optimisation 1(4/2010), 330–343 (2010), doi:10.1504/IJMMNO.2010.03543
7. Saida, I.B., Nadjet, K., Omar, B.: A New Algorithm for Data Clustering Based on Cuckoo Search Optimization. In: Pan, J.-S., Krömer, P., Snášel, V. (eds.) Genetic and Evolutionary Computing. AISC, vol. 238, pp. 55–64. Springer, Heidelberg (2014), doi:10.1007/978-3-319-01796-9_6.
8. Civicioglu, P., Besdok, E.: A Conceptual Comparison of the Cuckoo-search, Particle Swarm Optimization, Differential Evolution and Artificial Bee Colony Algorithms. Artificial Intelligence Review 39(4), 315–346 (2013), doi:10.1007/s10462-011-9276-0
9. Civicioglu, P., Besdok, E.: Comparative Analysis of the Cuckoo Search Algorithm. In: Yang, X.-S. (ed.) Cuckoo Search and Firefly Algorithm. SCI, vol. 516, pp. 85–113. Springer, Heidelberg (2014), doi:10.1007/978-3-319-02141-6_5.
10. Saida, I.B., Kamel, N., Omar, B.: A New Hybrid Algorithm for Document Clustering Based on Cuckoo Search and K-means. In: Herawan, T., Ghazali, R., Deris, M.M. (eds.) Recent Advances on Soft Computing and Data Mining SCDM 2014. AISC, vol. 287, pp. 59–68. Springer, Heidelberg (2014), doi:10.1007/978-3-319-07692-8_6.
11. Kamel, N., Ouchen, I., Baali, K.: A Sampling-PSO-K-means Algorithm for Document Clustering. In: Pan, J.-S., Krömer, P., Snášel, V. (eds.) Genetic and Evolutionary Computing. AISC, vol. 238, pp. 45–54. Springer, Heidelberg (2014), doi:10.1007/978-3-319-01796-9_5
12. Cagnina, L., Errecalde, I.M.: An Efficient Particle Swarm Optimization Approach to Cluster Short Texts. Information Sciences 265, 36–49 (2014), doi:10.1016/j.ins.2013.12.010

13. Song, W., Zhen Liang, J., Cheol Park, S.: Fuzzy Control GA with a Novel Hybrid Semantic Similarity Strategy for Text Clustering. Information Sciences 273, 156–170 (2014), doi:10.1016/j.ins.2014.03.024
14. Azaryuon, K., Fakhar, B.: A Novel Document Clustering Algorithm Based on Ant Colony Optimization Algorithm. Journal of Mathematics and Computer Science 7, 171–180 (2013)
15. Devi, S.S., Shanmugam, A.: Hybridization of K-means and Harmony Search Method for Text Clustering Using Concept Factorization. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 3(8) (August 2014)
16. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18(11), 613–620 (1975), doi:10.1145/361219.361220.
17. Huang, A.: Similarity Measures for Text Document Clustering. In: NZCSRSC 2008, Christchurch, New Zealand (April 2008)
18. Zhao, Y., Karypis, G.: Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. Machine Learning 55, 311–331 (2004), Kluwer Academic Publishers. Manufactured in The Netherlands
19. Xing, B., Gao, W.-J.: Cuckoo Inspired Algorithms. In: Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms. ISRL, vol. 62, Part II, Ch. 7, pp. 105–121. Springer International Publishing, Switzerland (2014), doi:10.1007/978-3-319-03404-1_7
20. Classic3 and Classic4 DataSets, Tunali, Volkan, http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/
21. Text retrival conference TREC, http://trec.nist.gov/

Information Technology: Requirement Engineering

Supporting Legal Requirements in the Design of Public Processes

Amina Cherouana(✉) and Latifa Mahdaoui

University of Sciences and Technology Houari Boumediene (USTHB), Algiers, Algeria
{acherouana,lmahdaoui}@usthb.dz

Abstract. Nowadays, business processes have become an ubiquitous part in public institutions, and the success of an e-government system depends largely on their effectiveness. However, despite the large number of techniques and technologies that are successfully used in the private sector, these cannot be transferred directly to public institutions without taking into account the strongly hierarchical nature and the rigorous legal basis on which public processes are based. This work presents an approach allowing the consideration of the legal requirements during the public processes design. Its main particularity is that these requirements are encapsulated using a legal features model supporting a formal semantic. This one prevents the violation of legal requirements and ensures that the processes evolution will in compliance with them.

Keywords: E-government · Information and Communication Technologies (ICT) · Business Process Management (BPM) · Public Process Design · Legal Requirements

1 Introduction

E-government is a phenomenon of an era in which e-business is becoming vital in both the private and the public sector. It is composed of a set of administrative processes (considered as business processes) whose mission is to serve citizens or businesses.

Indeed, the concept of business process has become an ubiquitous part in public institutions, and the success of an e-government system depends largely on their effectiveness. Consequently, the enormous and the spectacular benefits achieved in the industry and the private sector through the adoption of Business Process Management (BPM) haven't been without impact on public institutions. Let's note that the BPM is a process-centric approach which includes concepts, methods and technologies to support the design, administration, configuration, enactment, and analysis of business processes.

However, despite the large number of techniques and technologies that are successfully used in the private sector, these cannot be transferred directly to public institutions without taking into account the strongly hierarchical nature and the rigorous legal basis

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 231–242, 2015.
DOI: 10.1007/978-3-319-19578-0_19

on which public processes are based. The Government Process Management (GPM) is the thinking that derives from the application of BPM for public processes [18][19][20]. The process models in a such context are characterized by a set of rules, principles and specific models, collectively here referred to as legal requirements.

This work focuses on the design of public processes. Hence, the main problem in this context is to say: 'how to ensure that the designed public process models are on conformity with the legal framework governing public institutions?'. Under this issue, this paper proposes an approach allowing the consideration of the legal requirements during the public processes design. Let's specify that the legal requirements are mentioned in the law and the set of legal texts which constitute a source of valuable and incontrovertible knowledge.

The main particularity of this approach is that the legal requirements are encapsulated using a legal features model supporting a formal semantic. This semantic prevents the violation of legal requirements and ensures that the processes evolution is in compliance with them. In addition, the legal features model constitutes the core from which the first global models of public processes will be derived. These ones are, then, enriched with organizational aspects undescribed in the law and specific to each institution. Let's note that the legal features model is implemented using the Ontology Web Language (OWL) based on the Description Logics (DL) and the first, as well as the final, models of public processes are generated using the Workflow Nets formalism (WfN).

The remaining of this paper is structured as follow:

- Several research works can be inscribed in the same category as this work and try to propose solutions for the consideration of the legal requirements. A classification of these works and our positioning regarding these ones are made in the second section.
- The presentations of the proposed approach, as well as the description of its component intentions and strategies are made in the third section.
- This approach was tested and validated with the cooperation of an annex of Algerian Fiscal Administration. An overview of the results is presented in the fourth section.

2 Related Works

The legal requirements are mentioned in the law which includes the set of decrees and legal texts that are associated to each public institution. These contain the set of components, management rules and instructions regarding a public administrative procedure [5][6][20]. They also regulate strictly how to create a certain output [4].

Consequently, the consideration of legal requirements characterizing public institutions has become a major preoccupation in several research works. A thorough study of these has allowed us to classify them into three different orientations as shown below:

Table 1. Related Works Classification

Orientation	**Description**	**Examples of related works**
Normative Studies	Refers to the works that focus on the description of legal requirements as a distinctive aspect of public institutions, and the demonstration of the importance of their consideration in the proposal of any IT solution	Ximeng & al., 2009 [1] Saarenpaa & al.,2003 [2] Lenk, 1997 [3]
Methodologies and tools	Refers to the works that attempt to develop appropriate methodologies, tools and techniques to the support of legal requirements. They also covers the works that propose specific conceptual and methodological frameworks	Ciaghi & al., 2011 [4] Schumacher & al., 2013 [5] Alpar & al., 2005 [6]
Compliance assessment and verification	Refers to the works whose purpose is the assessment of implementation results and the verification of the compliance degree with the legal aspect	Amboala & al., 2010 [7] Wastell & al., 2001 [8] Zuo & al., 2010 [9]

This work belongs to the second category. Among works explored in this category are that of [4] who combine the principle of Business Process Reengineering (BPR) with a goal-oriented framework in order to analyze and to model the law. The emerged processes are then visualized using a subset of UML diagrams. In the same sense, there exists the works of [5] which propose framework for extraction and feeding processes from legal texts. The framework applies pipes and filters architecture and uses NLP tools to perform information extraction steps. A third example is that of [6] who focus on the problem of legal requirements modeling using the EPC language (Event-Driven Process Chain). They propose, then, an extension of the graphical notations of this language very responded in the business field.

The main particularity of the solution developed comes to the use of a formal semantic for a legal requirements support. In the following, a detailed description of the proposed solution and its intentions is made.

3 Approach Description

As mentioned above, the main concern of this work is the consideration of the legal requirements during the design of public processes. Therefore, the proposed solution consists, firstly, to encapsulate them into a legal features model supporting a formal semantic. This latter is represented using an ontological framework devoted to the semantic conception and implementation of public processes. It allows preventing the violation of legal requirements and ensuring that the processes evolution will in compliance with them.

The resulting legal features model constitutes the core from which the first global models of public processes must be derived. At this level, passage rules have been defined and implemented to ensure the automatic passage. These models are, then, enriched with organizational aspects undescribed in the law and specific to each institution in order to generate public process models. The figure.1 shows in detail the different intentions and strategies of the proposed approach.

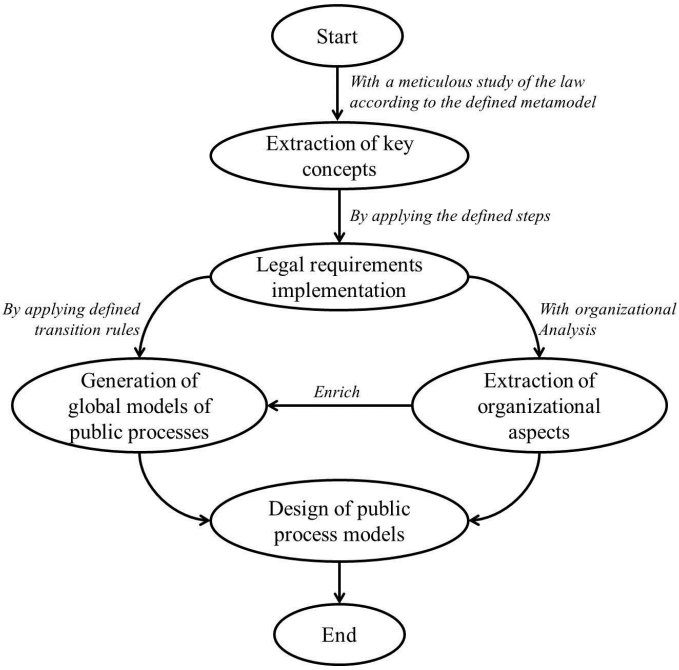


Fig. 1. Approach for Public Process Design

We use the MAP model [25] in order to represent clearly the approach phases, as well as the interrelations between them. The map is represented using an oriented and a labeled graph. The nodes represent intentions and the links represent strategies.

3.1 Extraction of Key Concepts

The first intention to achieve is the extraction of the key concepts which will be used in the legal requirements implementation intention. Hence, the starting point of this approach is a "*meticulous study of the law*" governing the targeted institution.

Several types of law exist, therefore it is important to operate a selection procedure and keep only those which provide information and knowledge that can be instantiated in the process (e.g. executive decrees and procedural decrees).

We have developed a law meta-model below (Figure 2) to describe the main concepts to be extracted. It covers all components that must be addressed in public processes design. Let's note that a key concept must not be questioned during the design process: it is necessary but not sufficient. A law is structured as several articles. It represents the primordial source providing the key concepts grouped in the following dimensions:

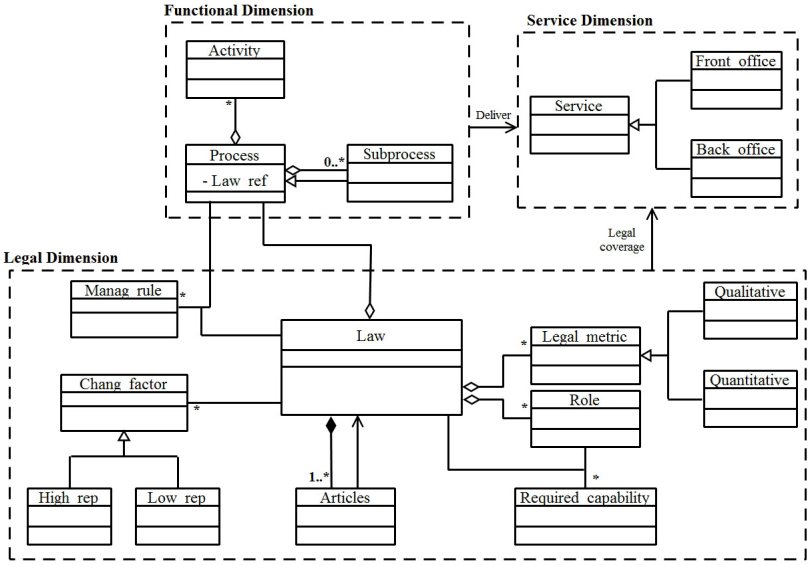


Fig. 2. Law Meta-Model for Public Process Design

3.2 Legal Requirements Implementation

Once the key concepts extraction was established, we pass to the implementation of the legal requirements governing the public processes. The main objective is to encapsulate the legal requirements through assets serving as the basis for the prevention of their violation. This one includes the definition of process parts, the structural relationships, as well as the description of dependencies between processes. The three "*defined steps*" to achieve this intention are shown in the figure 3:

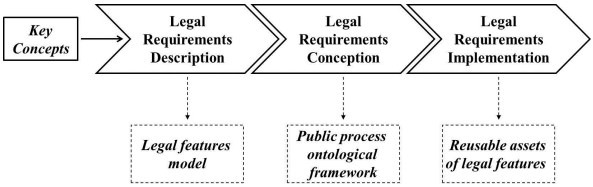


Fig. 3. Defined Steps for Legal Requirements Implementation

Legal Requirements Description
The objective of this phase is the generation of legal features model encapsulating the legal requirements. This phase uses as input the key concepts derived from the previous intention.

We adopt the feature model of Feature-Oriented Domain Analysis method [10]. This model is an explicit representation as a tree where nodes constitute the set of characteristics and the arcs specify the relationship between them. Let's specify that some features may have variations to choose and which will be resolved using the description logic (DL) during the implementation of legal requirements.

Legal Requirements Conception

It essentially comprises the construction of public processes ontology through a specific ontological framework. Indeed, we have defined a specific ontological framework for the semantic representation of public processes based on the legal features model. It is composed of two levels: (1) ontological framework associated to a public process, and (2) ontological framework associated to a public activity.

We have used the method of Uschold and King [12] who propose a method for enterprise's ontology construction. This latter is a two-level ontology, where the high-level is used to describe the domain concepts which, for their part, are placed in the second level [13][14]. We have selected the following corpus to describe the high-level

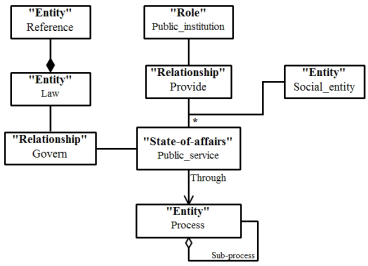


Fig. 4. Ontological Framework for a Public Process

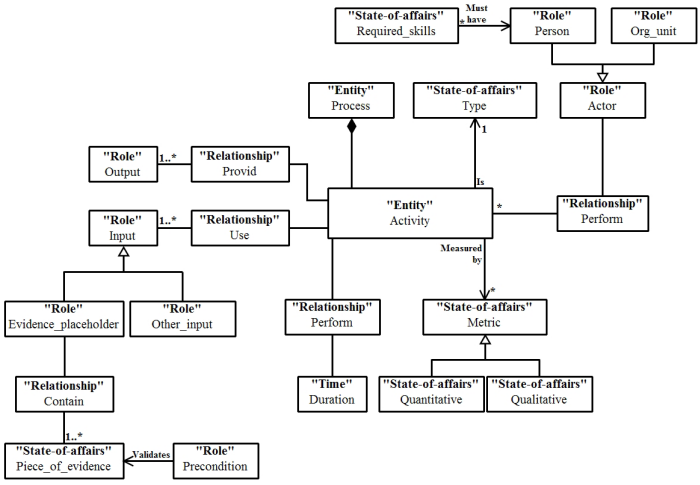


Fig. 5. Ontological Framework for a Public Activity

of our ontology: Entity, Relationship, Role, State-of-affairs and Time. These concepts are required to model any public process.

Legal Requirements Implementation
Legal requirements implementation is made with Ontology Web Language (OWL) based on the Description Logics (DL). Its main objective is the creation of assets that provide a basis from which the public process models will be derived.

This step includes also the configuration of the processes from the static variations points using a set of description logic axioms. This one favors the reuse of assets [15][16], prevent the violation of dependencies between variants of the features model by treating characteristics as components and dependencies as constraints [17].

3.3 Generation of Global Models of Public Processes

The purpose of this intention is the generation of the first global models as a Workflow Nets. Thus, a transformation rules allowing the passage from legal features model to Workflow Nets has been defined. For space reasons, these rules can't be presented in this paper. However, an example of the resulting global model will be presented in the relevant section in the case study.

This phase includes also the definition of execution order and the configuration of dynamic variations points. At this level, it is not possible to add behavior that has not been modeled beforehand and therefore not described in the law. Thus, all possible behaviors described in the law must appear in the resulting model. To solve this problem, we have done recourse to the approach proposed by Gottschalk & al. These authors have developed a configuration approach which is based on the restriction of the behavior for the Workflow Nets [22][23].

3.4 Extraction of Organizational Aspects

This intention covers mainly the: (1) identification of quick gains by identifying the flow in accordance with law, (2) collecting metrics of the current processes which allows, on one hand, to enrich those described in the law and produce an analytical view of the organization, and on the other hand, to establish a baseline for measurement and improvement of future processes, (3) extraction of actors with their appropriate skills in order to identify those able to occupy the roles extracted from the law, and to identify the need to improve capacity or to define new roles [20][21].

3.5 Design of Public Process Models

This is the intention where the public process models conform to the law are delivered. It comprises the necessary steps to transform the global dynamic models to the implementable models. It is during this phase where the integration of the organizational aspect is made. This last consists to define new activities/additional processes, new options and alternatives for processes within the project. It also includes the description of created or redefined jobs, the assignments of roles according to their capacity, as well as defining of business and managerial personnel with their job objectives. The manner in which their performance will be measured and managed is also changed or developed.

4 Case Study: Algerian Fiscal Administration

The Algerian fiscal administration is a public institution responsible for establishing the tax base, its perception and its control. It belongs to the category G2B (Government to Business), which imposed the study of the different stakeholders, as well as their rights and obligations towards the fiscal administration. Three categories of taxpayers are distinguished: (1) physical person, (2) capital company, and (3) foreign company which is divided to those installed and others not installed. For the test and the validation of this approach we have cooperate with the local Annex of Algiers.

4.1 Extraction of Key Concepts

A set of decrees (between executive and procedural) were selected to analyze and to extract legal requirements related to the tax regime on which taxpayers are subject (ex. Decree N°. 96-31, Decree N°. 08-98, Decree N°. 01-353, etc) [26]. The analysis of these documents has allowed identifying more than twenty processes each having a set of associated key concepts. These are conforming to the key concepts described in the law meta-model (Figure 2).

4.2 Legal Requirements Implementation

Legal Requirements Description
A fragment of the resulting legal features model related to the Algerian Fiscal Administration is shown in the following figure:

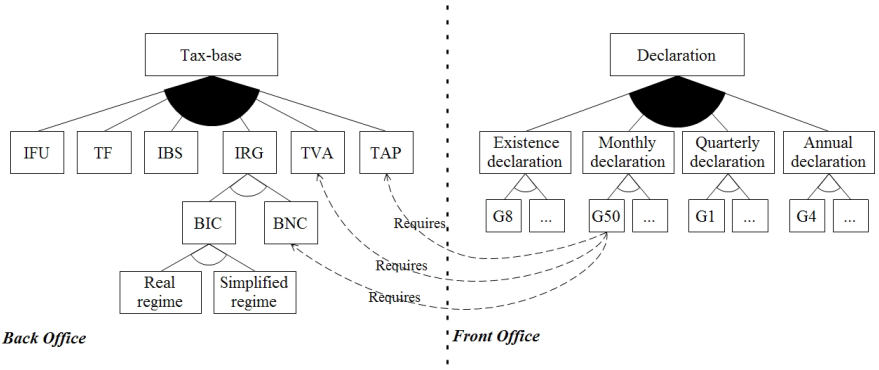


Fig. 6. Fragment of Legal Features Model of the Algerian Fiscal Administration

The back office represents the set of internal processes of the fiscal administration. For example, the global process "Tax-base", is composed of all taxable procedures described in the law (IFU, TF. . .). The front office represents the set of provided services to the different taxpayers. For example, "Declaration" is composed of all statements that the concerned must declare (Existence declaration, Monthly declaration...).

Legal Requirements Conception

The purpose of this phase is to build public processes ontology. It is made from the high-level ontology (Figure 4 and Figure 5). Let's note that each public process and its component activities must be designed and then implemented. An example of public process ontology (the Monthly declaration process) is shown in the figure7.

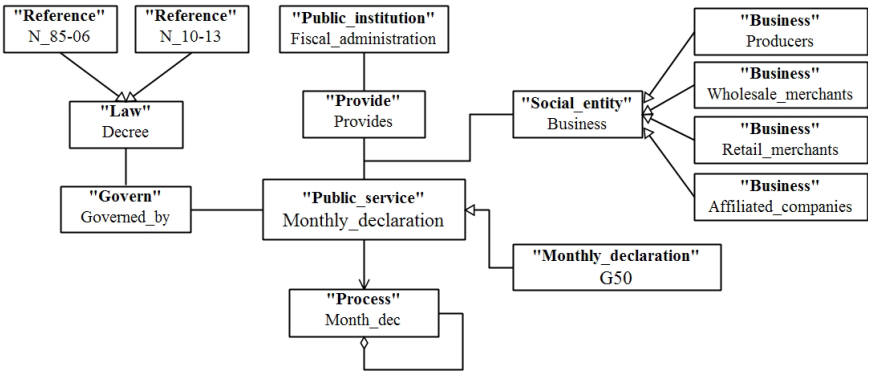


Fig. 7. Ontology Overview of the "Monthly_declaration" Process

Legal Requirements Implementation

The implementation of the legal requirements starts with the implementation of the high-level ontology with its components concepts and dependencies between them. The latter is, subsequently, imported to create the assets of different legal features. Remember that a set of axioms is also implemented in this phase. In addition, the assets consistency and the concepts classification and positioning have been checked using a specific reasoner, before their use in the next intention.

4.3 Generation of Global Models of Public Processes

This phase must be initiated by the generation of the first global models of public processes by applying defined rules, and defining the execution order of the extracted components. As example, the application of the defined rules on the back-office gives the following model:

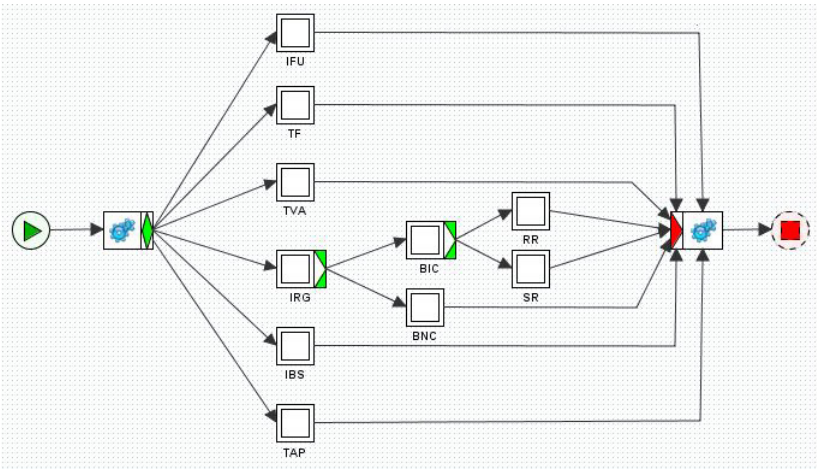


Fig. 8. Example of a Global Public Process Model

Let's recall that the resolution from the dynamic variation points was also made according to the mentioned approach [22][23].

4.4 Extraction of Organizational Aspects

This phase begins with the representation of current operational processes with BPMN notation. BPMN is located at the analysis level. It was introduced to provide a graphical notation easy to understand. These current models are, subsequently, analyzed and confronted with implemented assets.

This analysis allowed extracting several flows, activities, qualitative/ quantitative operational metrics and identifying needs to define new alternatives in the next phase. For extracted roles, establishing the matrix of capabilities [24] has provided useful information on current and future skills needs.

4.5 Design of Public Process Models

An overview of public process model "IFU" is shown in Figure 9. The development of this model is made by integrating the organizational aspect delimited by the constraints

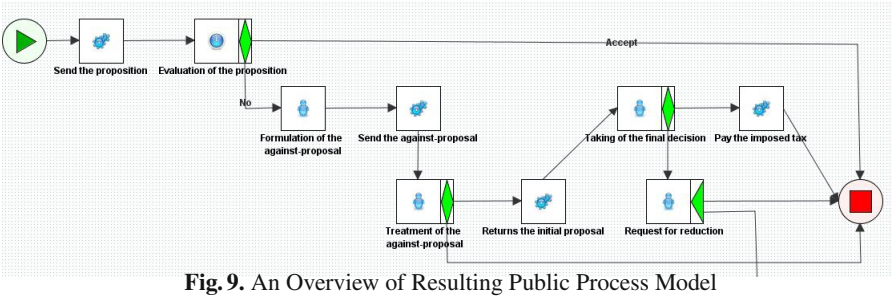


Fig. 9. An Overview of Resulting Public Process Model

of the implemented assets (ex. the sub-processes must be triggered by the tax administration, the taxpayer has 30 days to express its decision, etc.) and other issues from the previous phase (ex. time allowed for the tax inspector in order to treat against-proposal).

5 Conclusion

The objective of efficiency and effectiveness improvement of the e-government is a primordial problem. The processes of such system must obey to certain requirements of process models described in the law and the set of legal texts. For this fact, the internal processes are partially ruled and governed by a legal framework.

We have focused in this work to propose a design approach allowing the consideration and the support of the legal requirements governing a given public institution. The main particularity of this approach is that the legal requirements are encapsulated using a features model supporting a formal semantic. This last is represented using an ontological framework devoted to the semantic conception and implementation of public processes.

Several aspects can also be developed in order to evolve this approach. We focus now on the first intention and we try to develop a cooperative platform for the meticulous study of the law strategy in accordance with the law meta-model presented previously.

References

1. Ximeng, L.: Research on E-government Initiatives and Enabling IT, PhD thesis, University of Hong Kong (2009)
2. Saarenpää, A.: A Legal Framework for *E*-Government. In: Traunmüller, R. (ed.) EGOV 2003. LNCS, vol. 2739, pp. 377–384. Springer, Heidelberg (2003)
3. Lenk, K.: Business process reengineering in the public sector: opportunities and risks. In: Beyond BPR in Public Administration: An Institutional Transformation in an Information Age, pp. 151–165. IOS Press (1997)
4. Ciaghi, A., et al.: Villa_orita, Improving Public Administrations via Law Modeling and BPR. In: AFRICOM (2011)
5. Schumacher, P., Minor, M., Schulte-Zurhausen, E.: Extracting and enriching workflows from text. In: 14th International Conference on Information Reuse and Integration (IRI). IEEE (August 2013)
6. Alpar, P., Olbrich, S.: Legal requirements and modelling of processes in e-government. Electronic Journal of e-Government 3(3), 107–116 (2005)
7. Amboala, T., Japang, M., Likoh, J., Yuszreen, M.: Business Process Reengineering In Labuan Fire Services Operations: A Case Study. Labuan e-Journal of Muamalat and Society LJMS – Special Issue 4(2010), 14–25 (2010)
8. Graham Wastell, D., Kawalek, P., Willetts, M.: Designing alignment and improvising change: Experiences in the public sector using the SPRINT methodology. In: ECIS 2001, pp. 1125–1136 (2001)
9. Zuo, L., Liu, Y.: Notice of Retraction Organizational Change Pattern Based on Business Process Reengineering. In: International Conference on E-Business and E-Government (2010)
10. Kang, K., Cohen, S., Hess, J., Novak, W., Peterson, A.: Feature-Oriented Domain Analysis (FODA) Feasibility Study, Technical Report CMU/SEI-90-TR-21, SEI. Carnegie Mellon University, Pittsburgh, Pennsylvania (1990)

11. Thum, T., Kastnery, K., Erdwegy, S., Siegmund, N.: Abstract Features in Feature Modeling. In: 15th International Software Product Line Conference (2011)
12. Uschold, M., King, M., Moralee, S., Zorgios, Y.: The enterprise ontology. Knowledge Engineering Review 13(1), 31–90 (1996)
13. Penicina, L.: Choosing a BPMN 2.0 Compatible Upper Ontology. In: eKNOW 2013: The Fifth International Conference on Information, Process, and Knowledge Management (2013)
14. Semy, S.K., Pulvermacher, M.K., Obrst, L.J.: Toward the Use of an Upper Ontology for U. S. Government and U. S. Military Domains: An Evaluation, Corporate Head-quarters Bedford, Massachusetts (2004)
15. Huang, Y., Key, S.: Ontology-Based Configuration for Service-Based Business Process Model. In: IEEE International Conference on Services Computing (SCC) (2013)
16. Döhring, M., Reijers, A., Smirnov, S.: Configuration vs. adaptation for business process variant maintenance: An empirical study. Information System Journal 39, 108–133 (2013)
17. Mafazi, S., Mayer, W., Grossmann, G., Stumptner, M.: A Knowledge-based Approach to the Configuration of Business Process Model Abstractions. Journal of Knowledge-Based Configuration- Survey and Future Directions 15, 47–66 (2012)
18. Zhang, N., Hou, X.: Government Process Management under electronic government and its application. In: International Conference on E-Business and E-Government (ICEE), pp. 1–4 (2011)
19. Xuefang, X.: Study of government information construction based on BPR. In: International Colloquium on Computing, Communication, Control, and Management, CCCM 2009, vol. 1, pp. 318–320 (2009)
20. Cherouana, A., Mahdaoui, L.: Towards a methodological framework for the Government Process Management. In: International Conference on Research Challenges in Information Science (RCIS), Valencia, Spain (2012)
21. Cherouana, A., Mahdaoui, L.: Study of OSSAD applicability in a GPM framework. In: International Conference on Electronic Governance (ICEGOV), Seoul, Republic of Korea (2013)
22. Gottschalk, F., Van der Aalst, W., Jansen-Vullers, M., Marcello, L.R.: Configurable Workflow Models. Int. J. Cooperative Inf. Syst. 17(2), 177–221 (2008)
23. Gottschalk, F., Wagemakers, T.A.C., Jansen-Vullers, M.H., van der Aalst, W.M.P., La Rosa, M.: Configurable Process Models: Experiences from a Municipality Case Study. In: van Eck, P., Gordijn, J., Wieringa, R. (eds.) CAiSE 2009. LNCS, vol. 5565, pp. 486–500. Springer, Heidelberg (2009)
24. Jeston, J., Nelis, J.: Manage by Process - A Roadmap to Sustainable Business Process Management. Published by Elsevier Ltd., (2008)
25. Rolland, C., Prakash, N., Benjamin, A.: A Multi-Model view of Process Modelling. Requirement Engineering 4, 169–187 (1999)
26. Algerian Official Journal – JORA, www.joradp.dz/

Requirement Analysis in Data Warehouses to Support External Information

Mohamed Lamine Chouder(✉), Rachid Chalal, and Waffa Setra

LMCS (Laboratoire de Methodes de Conception de Systemes),
ESI (Ecole nationale Superieure d'Informatique), Algiers, Algeria
{m_chouder,r_chalal,w_setra}@esi.dz

Abstract. In strategic decision-making, the decision maker needs to exploit the strategic information provided by decision support systems (DSS) and the strategic external information emanating from the enterprise business environment. The data warehouse (DW) is the main component of a data-driven DSS. In the field of DW design, many approaches exist but ignore external information and focus only on internal information coming from the operational sources. The existing approaches do not provide any instrument to take into account external information. In this paper, our objective is to introduce two models that will be employed in our approach: the requirement model and the environment model. These models are the basis of our DW design approach that supports external information. To evaluate the requirement model, we will illustrate with an example how to obtain external information useful for decision-making.

Keywords: Data warehouse · Design · Requirement analysis · External information · Business environment

1 Introduction

A data warehouse (DW) that supports external information is a knowledge source for strategic decision-making. It provides historical information about the enterprise business environment. External information is the strategic information useful for decision-making, about competitors, customers, markets, suppliers, products. Unfortunately, this type of information is informal, heterogeneous and unstructured, which makes the process of developing a DW that satisfies decision-makers needs a difficult and a complex task. For many years, it is widely accepted that the basis for designing a DW is multidimensional (MD) modeling [1, 2]. Today, the MD form is natural to decision makers, by means of its structure composed of analysis measures and dimensions that represent the context for analysis.

In the literature, two different categories of DW design approaches exist: data-driven and requirement-driven. The former starts from operational sources to define the MD model of the DW [3, 4]. The latter tries to identify the requirements to build the DW and define its contents [5]. These approaches collect requirements through different orientations: users, processes, and goals and using different techniques

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 243–253, 2015.
DOI: 10.1007/978-3-319-19578-0_20

(See section 2). In this work, our interest is concentrated on requirement-driven, mostly goal-oriented, DW design approaches for two reasons: (i) the strategic goals of the organization in the business environment are considered the main resource to identify external information requirements; (ii) the structure of the external information source, which will aliment the DW, is not defined, unlike in operational sources, so it must be defined. We argue that in the field of DW design, the existing approaches focus only on internal information coming from operational sources. These approaches raise the importance of external information, but ignore it and do not provide any instrument to support it.

To answer this, we propose a goal-oriented requirement analysis approach in DWs to support external information. This approach can be used to build a DW that contains strategic information useful for decision-making. In this paper, the models that will be employed in our approach are detailed: the requirement model and the environment model. The former is an improved requirement model from the model proposed in [6]. This model aims to identify the strategic decision-making needs for external information. The latter represents useful information for strategic decision-making about some environment elements (competitor, customer, market, product, and supplier). In a future work, a set of processes will be defined to show how to use the models described in this paper for defining external information requirements and the underlying MD model. To evaluate our proposal, the requirement model will be illustrated using an example: the strategic goal "Increase market share". To do this, a set of guidelines are defined to show how to obtain external information useful for decision-making.

The remainder of this paper is organized as follows. Section 2 discusses related work. In Section 3, the models employed in our approach are described. Section 4 represents the illustration of the requirement model with an example. Finally, Section 5 points out the conclusion and future work.

2 Related Work

In the last decade, various DW design approaches were proposed to define requirements using different techniques. For example: Business process models in [7], Goal-Question-Metric approach in [8], use cases in [9], best practices in [10], traditional requirements engineering (RE) process in [11], Decision processes analysis in [5], Map goal model in [12], GDI model in [13], extended Tropos in [14], extended i* framework in [15], etc... Due to lack of space, in this section we give a brief description of the most relevant requirement-driven, mostly goal-oriented, DW design approaches.

Starting with [13], DW requirements are determined in the broader context of the goals and objectives of an organization. At first, in an organizational perspective, requirements are grouped into several levels of abstraction using the Goal-Decision-Information (GDI) model. It starts by determining goals, then the decisions that influence the satisfaction of these goals. Finally, the information needed to make decisions is identified. At second in a technical point of view, information scenarios are applied

for each decision, to define DW contents and their proprieties. This approach shares similarities with ours in abstraction levels and because it deals with decisional goals, however, it does not consider external information.

[14] presented GRAND, a goal-oriented approach, which has extended the early phase of Tropos [16] to the requirements engineering of DWs. Tropos is an agent-oriented method, which is a variant of i* [17]. In requirement analysis, the stakeholder's dependencies are represented in an actor diagram. Then, two perspectives are adopted, organizational and decisional. In the former, facts and attributes are identified and associated with goals of different actors. In the latter, each fact is related to their dimensions and a set of measures is found out and associated with facts. In conceptual modeling, this approach can be either employed, within requirement-driven or mixed requirement/data-driven, to specify the conceptual MD model. This approach share similarities with ours in goal reasoning. However, it does not consider decisional goals unlike [13] and does not provide any instrument to take into account external information.

In [15] another goal-oriented approach is proposed based on the i* framework [17]. At first, the strategic, decisional and informational goals are identified through interviews. The information requirements (tasks and resources) are obtained from the informational goals of different actors using two i* models: Strategic Dependency (SD) model and Strategic Rational (SR) model. This latter is applied for the DW actor to define the rational model, which will give rise to the design of a conceptual MD model using a UML profile [18]. In this approach, organizational modeling unlike [14] and external information are not supported.

In recent years, many researchers focus on understanding the business context in which the DW will be implemented. In [19] an extended version of the work in [15] was proposed to align DW requirements with the business strategy. This approach considered the business strategy using VMOST (vision, mission, objective, strategy, and tactics) and the business motivation model (BMM) to align DW goals and the organization strategy. In the same direction, another effort has been made by [20] to align the i* concepts for requirement analysis in DWs with the business strategy model proposed in the business intelligence model (BIM) [21].

BIM is a business modeling language that offers many diagrams to help business users make sense of data manipulated in business intelligence systems. Different reasoning techniques about goals, goal influences, situations and indicators are used to define a complete business strategy plan [22]. The interesting thing for us in this model is that external and internal situations, that influence the fulfillment of a goal, are identified. Then, one or many external indicators are associated with an external situation (e.g. number of competitors). However, the authors do not indicate how to identify these indicators and how to represent them in a data perspective (dimensions and measures). In addition, this model is used to shape the organizational strategy, not to build a DW.

Although, the formalisms used in the presented approaches for requirement analysis step are different, their expressivity is very close, and show that a core of common information has been identified [23]. In the next section, we will present an outline of our approach for requirement analysis in DWs to support external information, which is not supported in the above-presented approaches.

3 Requirement Analysis in DWs to Support External Information

In strategic decision-making, the decision maker needs to exploit mainly the information emanating from the enterprise business environment. This information is qualified as external information. Since DW design approaches try to provide organizations with information to support decision-making. A DW design approach should be defined to support external information useful for decision-making. Therefore, better decisions will be taken and strategic goals are achieved.

In our work, the strategic objectives that the organization must achieve are considered the strategic goals to identify external information. This latter is the information required to support decision-making that must be provided by the DW. A number of proposals for requirement analysis in DWs have been made as seen in section 2 with focus on information coming from operational sources. Many efforts also have been made to define the system development lifecycle (SDLC) for DW development. Some approaches take the ER diagram or the database scheme of operational sources as an input to their DW requirement analysis stage [4]. In [13], the authors argue that DW development must be rooted by the set of goals and decisions interesting the organization rather than the schema of operational sources.

In our approach, we propose a hybrid SDLC (See table 1) where the DW schema and the source schema[1] are defined at the same time. In requirement analysis, two models are adopted: a requirement model and an environment model. The former is used to identify external information required for decision-making. The latter is a UML class diagram, which will be used in requirement analysis to derive information that must be provided by the DW represented in the form of dimensions and measures. In addition, in conceptual design, this model will give rise to the design of the DW multidimensional schema and the schema of the data source that will aliment the DW.

Table 1. The proposed SDLC

Stage	Output
Requirement analysis	Requirement diagram, Specific environment model
Conceptual design	DW multidimensional schema, Source conceptual schema
Logical design	DW logical schema, Source logical schema
Physical design	DW physical schema, Source physical schema

In the following subsections, the models that will be employed in our approach are described in detail. The manner of using the described models together in requirement analysis is not addressed in this paper. This aspect will be extended in a future work.

[1] The source schema is the structure of the data source that will aliment the DW in contrast with operational sources

3.1 The Requirement Model

The requirement model is an improvement to the strategic goal model proposed in a previous work [6]. In our model, a strategic goal is considered as an objective that is to be met by the organization at the strategic level. From the implementation process of a strategic goal, a set of goals are derived [6]. Once a goal is defined, it either needs means to concretize it or decisions to realize it.

As shown in Fig.1 a strategic goal consists of a set of goals, which makes a goal hierarchy. A goal can be either qualified as an operational goal or a decisional goal. The latter is a long-term goal in the strategic/tactical level of the organization that needs decisions to realize it. Whereas an operational goal is an objective that can be met by a transactional information system, which is concretized by realization and control means.

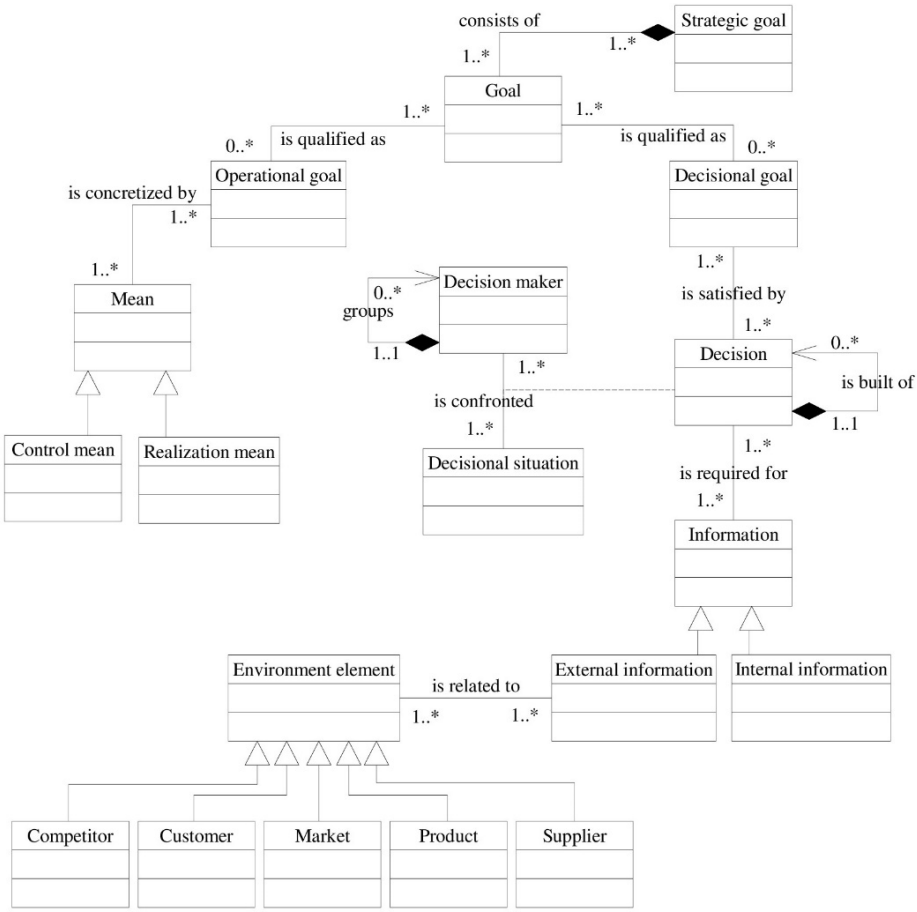


Fig. 1. The requirement model

A decision is the intention to perform the actions that cause its implementation to fulfill decisional goals [13]. Decision-making is an activity that results in the selection of the decision to be implemented. While performing this activity, the decision maker is in a decisional situation where he/she requires the appropriate information to select the right decision among alternative decisions. As shown in Fig.1, a decision can be built out of other decisions as in [13]. A decision maker, which can be an individual or a group, is confronted to a decisional situation where decisions have to be taken. The association 'is satisfied by' between decisional goal and decision identifies the decisions which, when taken can lead to decisional goals satisfaction.

Knowledge necessary to take decisions is represented by the form of information. Fig.1 shows that there is an association 'is required for' between information and decision. This association identifies the information required to take a decision. The information can be external or internal. The internal information branch is out the scope of our interest. External information is the specification of data that will be stored in the DW. It is the information about environment elements (competitor, customer, market, product, and supplier). As shown in Fig.1, the association 'is related to' between external information and environment element identifies the information about the environment.

To offer a graphical support for requirement modeling, the notation of some elements from the i* framework [17] will be used. The graphical extended notation is summarized in Fig.2.

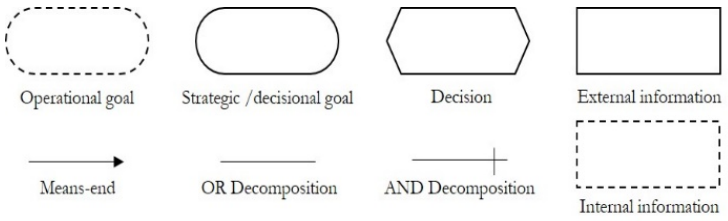


Fig. 2. The notation used in requirement modeling

3.2 The Environment Model

As seen in the previous subsection, information about the environment is necessary in strategic decision-making, which has an impact on the enterprise competitiveness. The business environment can be defined as the external factors that influence directly or immediately the enterprise. It is composed of two categories: (i) the macro-environment, which is the general environment that integrates political, economic, social, technological, legal… aspects; (ii) the micro-environment, that is our concern, is the close environment constituted with factors like customers, competitors, markets, products, and suppliers…

One of the largest used models in analyzing the environment is the Porter's five forces model [24]. It identifies five forces influencing the enterprise in a competitive environment: rivalry between competitors, threat of potential entrants, bargaining power of suppliers, bargaining power of customers, and threat of substitute

products/services. Taking into account these forces, we can assume that the five major elements of the enterprise environment are: competitor, customer, market, product, and supplier. The literature on environment analysis only proposes outlines and does not provide any structuration of information about the enterprise business environment [25]. Therefore, the proposal of an environment model, which will offer a generic view of this environment.

The first model produced by our research team is the competitor model [25] (See Fig.3). In this paper, only this model will be described in detail. Nevertheless, the other parts of the environment model have the same definition, as the competitor model, described in the next paragraph. The competitor model assembles informational bricks serving for the acquisition of information about the competitor. This model has been modeled based on the UML class diagram for many reasons: (i) it is largely used today, so it is familiar to designers; (ii) it permits to represent different points of view. In addition, to facilitate its navigation and use, the model has been conducted using the meta-modeling principle. Thus, it consists of two levels of modeling: meta-class level and class level.

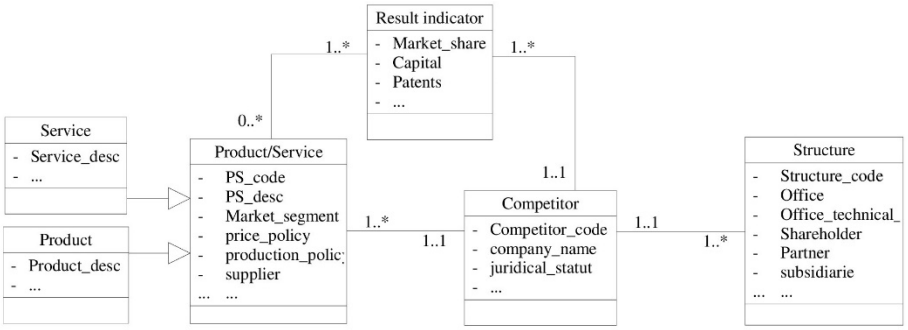


Fig. 3. Competitor meta-model [25]

As illustrated in Fig.3, the competitor meta-model regroups different meta-classes that are related to the competitor:

- Result indicator meta-class regroups financial, statistical, and commercial results about the competitor like capital, market share, patents…
- Structure meta-class represents: (i) information about the competitor identity, structural and organizational aspects, shareholders, partners…; (ii) information about the competitor-implemented strategies, techniques used for each enterprise domain: commercialization, distribution, projects funding, provision, research and development…
- Product meta-class is the main component in this model. It describes information about the competitor activity (products and services) and policies applied to products.

Fig.4 shows the product meta-model where the product class has many associations with other classes: campaign, customer opinion, market, supplier, price-policy, production-policy, promotion.

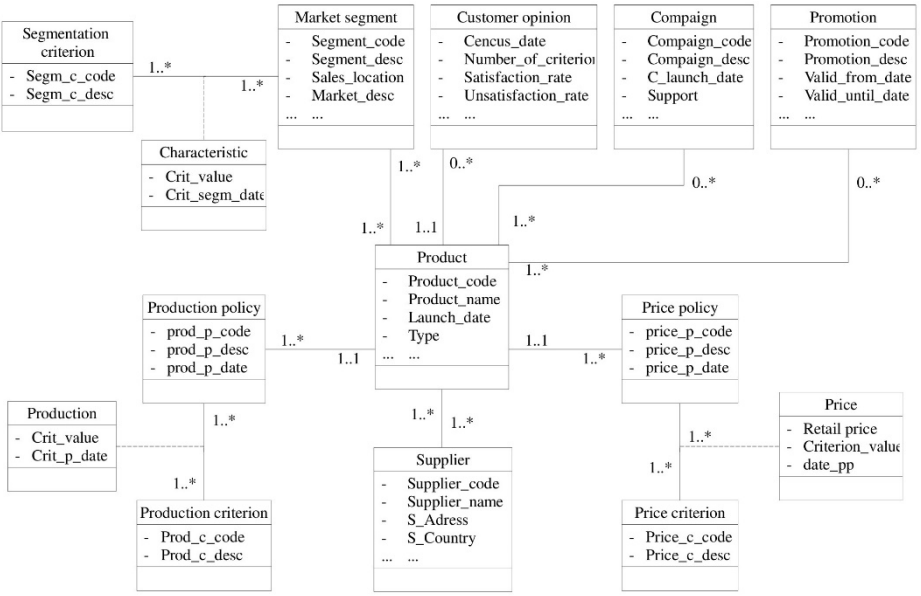


Fig. 4. The product meta-class of the competitor model [25]

In a future work, the environment model will be used in requirement analysis with the requirement model to shape the external information required in the form of dimensions and measures. In addition, it will be used in conceptual design to define the MD model and the schema of the data source that will aliment the DW. In the next section, the requirement model will be illustrated with an example and a set of guidelines are defined to show how to use it.

4 Sample Application of the Requirement Model

In this section, we propose a set of guidelines to show how to use the requirement model. These guidelines are used to demonstrate how to derive external information useful for decision-making from the strategic goal “Increase market share”. The outcome of requirement modeling is a requirement diagram.

Guideline 1. The process of implementing a strategic goal must be described. It will be analyzed to identify a set of goals relative to a strategic goal. See [6] for more details on how to do this task semi-automatically.

Guideline 2. The goals identified in the previous step are qualified as decisional goals or operational goals. Operational goals are excluded. The resulted hierarchy is modeled as a goal hierarchy using the goal notation and decomposition links.

Guideline 3. A set of decisions is identified for each decisional goal and linked to their respective goals with a means-end link. After that, complex decisions are decomposed. The decision hierarchy is modeled using the decision notation and decomposition links.

Guideline 4. For each decision, a decisional situation is identified. A decisional situation is the description of the decisional problem, which identifies the internal and external variables that influence the decision-maker when taking a decision. Therefore, it could define the nature of the information needed.

Guideline 5. From the decisions and decisional situations identified in the previous step, environment elements are identified. The external information needed to take a decision is defined in the form "information about environment elements". This external information is associated with a means-end link to each decision in the requirement diagram.

Example. Fig.5 shows the resulted requirement diagram of requirement modeling, after analyzing the strategic goal "Increase market share". This goal is built out of two decisional goals: "Increase sales" and "Retain customer loyalty". For the goal "Increase sales", two decisions are identified "Launch a new product" and "Open new sales channels". Moreover, for the goal "Retain customer loyalty", the decision "Improve quality of existing products" is identified.

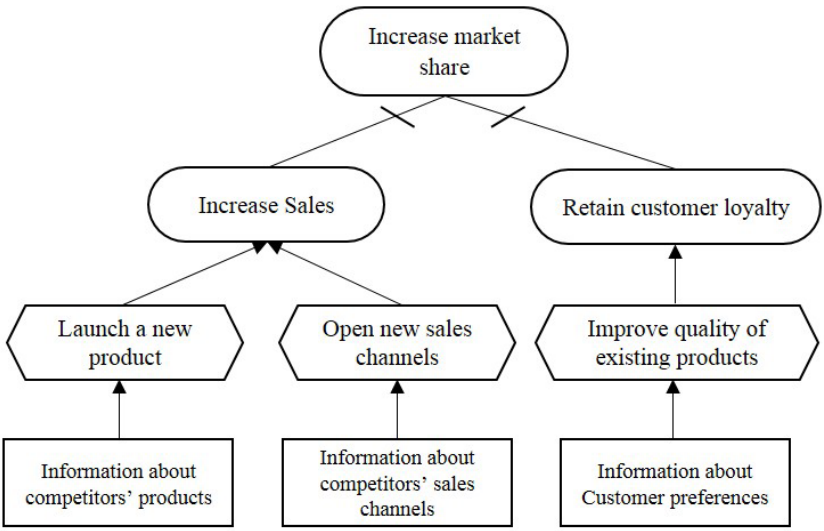


Fig. 5. Requirement diagram for the strategic goal "Increase market share"

For the decision “launch a new product”, two environment elements are identified product and competitor. Therefore, the external information “information about competitors’ products” is associated to the decision “launch a new product”. Identically “information about competitors’ sales channels” is associated with the decision “open new sales channels”. Furthermore, “information about customer preferences” is associated to the decision “improve quality of existing products”.

5 Conclusion

In this paper, we have presented the first step of a goal-oriented approach for requirement analysis in DWs to support external information. As existing approaches mainly focus on internal information coming from operational sources, our approach provides the means to build a DW that contains strategic information about the enterprise business environment. We were limited to the description of the models that will be employed in our approach: the requirement model and the environment model. The former is used to identify external information required for strategic decision-making. The latter represents information useful for decision-making about the enterprise business environment. Immediate planned future work involves defining a set of processes to show how to use these models to shape the external information required in the form of dimensions and measures. Then to obtain the underlying MD model and the structure of the data source that will aliment the DW.

References

1. Kimball, R., Ross, M.: The data warehouse lifecycle toolkit. John Wiley & Sons (2002)
2. Inmon, W.H.: Building the data warehouse. John wiley & sons (1996)
3. Golfarelli, M., Maio, D., Rizzi, S.: The dimensional fact model: a conceptual model for data warehouses. International Journal of Cooperative Information Systems 7, 215–247 (1998)
4. Hüsemann, B., Lechtenbörger, J., Vossen, G.: Conceptual data warehouse design. Universität Münster. Angewandte Mathematik und Informatik (2000)
5. Winter, R., Strauch, B.: A method for demand-driven information requirements analysis in data warehousing projects. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, p. 9. IEEE (2003)
6. Boukrara, A., Chalal, R.: Specification of useful information for the strategic decision support: risks-based approach. International Journal of Decision Sciences, Risk and Management 4, 276–293 (2012)
7. Böhnlein, M., Ulbrich-vom Ende, A.: Business process oriented development of data warehouse structures. In: Data Warehousing, pp. 3–21. Springer (2000)
8. Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S.: Designing data marts for data warehouses. ACM Transactions on Software Engineering and Methodology 10, 452–483 (2001)
9. Bruckner, R., List, B., Scheifer, J.: Developing requirements for data warehouse systems with use cases. In: AMCIS Proceedings, p. 66 (2001)

10. Schiefer, J., List, B., Bruckner, R.: A holistic approach for managing requirements of data warehouse systems. In: AMCIS Proceedings, p. 13 (2002)
11. Paim, F.R.S., de Castro, J.F.B.: DWARF: An approach for requirements definition and management of data warehouse systems. In: Proceedings of 11th IEEE International Requirements Engineering Conference, pp. 75–84. IEEE (2003)
12. Gam, I., Salinesi, C.: A requirement-driven approach for designing data warehouses. In: Proceedings of Requirements Engineering: Foundation for Software Quality (REFSQ) (2006)
13. Prakash, N., Gosain, A.: An approach to engineering the requirements of data warehouses. Requirements Engineering 13, 49–72 (2008)
14. Giorgini, P., Rizzi, S., Garzetti, M.: GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decision Support Systems 45, 4–21 (2008)
15. Mazón, J.-N., Pardillo, J., Trujillo, J.: A model-driven goal-oriented requirement engineering approach for data warehouses. In: Hainaut, J.-L., et al. (eds.) ER Workshops 2007. LNCS, vol. 4802, pp. 255–264. Springer, Heidelberg (2007)
16. Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., Mylopoulos, J.: Tropos: An agent-oriented software development methodology. Autonomous Agents and Multi-Agent Systems 8, 203–236 (2004)
17. Yu, E.S.-K.: Modelling strategic relationships for process reengineering. University of Toronto (1995)
18. Luján-Mora, S., Trujillo, J., Song, I.-Y.: A UML profile for multidimensional modeling in data warehouses. Data & Knowledge Engineering 59, 725–769 (2006)
19. Cravero Leal, A., Mazón, J.N., Trujillo, J.: A business-oriented approach to data warehouse development. Ingeniería e Investigación 33, 59–65 (2013)
20. Maté, A., Trujillo, J., Eric, S.: Aligning Data Warehouse Requirements with Business Goals. In: iStar, pp. 67–72 (2013)
21. Barone, D., Yu, E., Won, J., Jiang, L., Mylopoulos, J.: Enterprise modeling for business intelligence. In: van Bommel, P., Hoppenbrouwers, S., Overbeek, S., Proper, E., Barjis, J. (eds.) PoEM 2010. LNBIP, vol. 68, pp. 31–45. Springer, Heidelberg (2010)
22. Horkoff, J., Barone, D., Jiang, L., Yu, E., Amyot, D., Borgida, A., Mylopoulos, J.: Strategic business modeling: representation and reasoning. Softw. Syst. Model. 13, 1015–1041 (2014)
23. Golfarelli, M.: From User Requirements to Conceptual Design in Data Warehouse Design. In: Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction (2010)
24. Porter, M.E.: Competitive Strategy: Techniques for Analyzing Industries and Companies. Free Press (1980)
25. Chalal, R., Boukrara, A., Saddok, M., Guiri, S.: A model for the aquisition of competitor information for strategic decision support. In: ISKO-France (2013)

Engineering the Requirements of Data Warehouses: A Comparative Study of Goal-Oriented Approaches

Waffa Setra(✉), Rachid Chalal, and Mohamed Lamine Chouder

LMCS (Laboratoire de Methodes de Conception des Systemes)
ESI (Ecole nationale Superieure d'Informatique), Algiers, Algeria
{W_setra,r_chalal,m_chouder}@esi.dz

Abstract. There is a consensus that the requirements analysis phase in the development project of a data warehouse (DW) is of critical importance. It is equivalent to application of requirements engineering (RE) activities, to identify the useful information for decision-making, to be met by the DW. Many approaches has been proposed in this field. Our focus is on goal-oriented approaches which are requirement-driven DW design approaches. We are interested in investigating to what extent these approaches went well with respect to the RE process. Thus, theoretical foundations about RE are presented, including the classical RE process. After that, goal-oriented DW design approaches are described briefly; and evaluation criteria, supporting a comparative study of these approaches, are provided**.**

Keywords: Data warehouse · Goal-oriented approach · Requirements engineering process · Comparative study · Evaluation criteria

1 Introduction

In the last years, great interest has been shown in the field of Data warehouse (DW) design [1]. Indeed, many design approaches has been proposed in this field. These approaches are usually classified into two categories: data-driven and requirement-driven. The former also called supply-driven designs the DW starting from a detailed analysis of the data sources [1,2,3,4]. The user is not much involved in this category of approaches [5]. The latter also called demand-driven, attempts to identify the information requirements from business users [6,7,8,9]. We focus on the requirement-driven approaches.

Requirements analysis is the initial phase of DW design cycle [10]. It is equivalent to application of requirements engineering (RE) activities, to identify the useful information for decision-making, to be met by the DW. Requirement-driven DW design approaches define requirements through different orientations: process, user and goal. Process-oriented approaches [5], [11,12,13] analyze requirements by identifying the business processes of the organization. User-oriented approaches identify the target users and specify their individual needs to integrate them into a unified requirement model [14,13]. Goal-oriented approaches [8], [15,16,17,18,19]

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 254–265, 2015.
DOI: 10.1007/978-3-319-19578-0_21

identify goals and objectives of users that guide decisions at various levels of the organization. Most of requirement-driven DW design approaches are goal-oriented. Many authors recognize that these approaches provide a better definition of user requirements [15], [20], for two reasons: (i) the gathered requirements are validated by identifying conflicting goals; (ii) the different modelling alternatives to achieve a goal are provided [21]. However, in the beginning of a DW development project, identifying users' objectives and goals is a crucial step, where achieving the goals is an important indicator of the organization's activity [22].

RE is an important field dedicated to requirements definition. It is concerned by transforming users' expectations into agreed requirements through a well-defined process, called RE process. RE applied in the field of DW allows determining users' requirements. In this paper, our interest is on goal-oriented DW design approaches. Many approaches have been proposed in this field. As yet, there is no common strategy for these approaches [23]. Furthermore, we argue that the process of RE is not completely applied in this field. Besides, there is no common RE process for DWs. Indeed, if we consider that a structure of an approach is the set of its activities; the proposed approaches does not share the same structure. The purpose of this paper, is to extract the invariant steps from the classical RE process, in order to identify a set of criteria (see section 3.1), to allow evaluating the goal-oriented DW design approaches; in other words, to see what does each approach provide to support those criteria.

In this work, a general overview is shown, as well as a comparative study, of six famous goal-oriented approaches for DW design. The comparison highlights the evaluation criteria based on the classical RE process. The remainder of this paper is structured as follows: section 2 gives theoretical foundation of RE. In section 3, a brief description of goal-oriented approaches is given as well as the criteria to evaluate these approaches are defined in order to make a comparative analysis. Finally, Section 4 summarizes our work and presents our conclusions.

2 RE: Theoretical Foundation

2.1 What Is RE

The first definition of RE was given in the software engineering area [22]. It was qualified as visionary, referenced by many authors[24],[25] and specifies that: "requirements definition is a careful assessment of the needs that a system is to fulfil. It must say why a system is needed, based on current or foreseen conditions, which may be internal operations or an external market. It must say what system features will serve and satisfy this context. And it must say how the system is to be constructed" [26]. [26] stated that "requirements definition must encompass everything necessary to lay the groundwork for subsequent stages in system development".

Thus, RE must address firstly the 'why' dimension, justifying the existence of the system, which many authors translate to 'goal' or 'objective' [19], [27,28,29]. Then it addresses the 'what' dimension, specifying the system's functions to fulfil the goals

[27,28,29]. Besides, RE must take into account the 'how' dimension, by specifying the constraints to be applied on the system under consideration. Lamsweerde [27] has added the 'who' dimension, to address assigning responsibilities to humans, devices or softwares. While Zave [29] claimed that RE deals also with the evolution of the software's specifications over time. In RE, two types of requirements exist: Functional and non-functional. The former describes the functions to be performed by the system. The latter defines constraints on the way the functional requirement should be satisfied. A taxonomy of non-functional requirements can be found in [27].

The above-presented definition is taken from the software engineering field. In the literature, the aspects of RE engineering highlighted by the authors above were taken back in other fields: information systems [20], [30,31,32], DWs [17], [19], [33]. Since our interest is in the field of DWs, RE for DWs is detailed in section 3.

2.2 The RE Process

The RE process is composed of several activities highly intertwined. This property is observed at the different RE process models proposed [27], [34,35,36], while other authors [27], [34] , [36] affirm that the RE process is iterative and incremental. Two common concepts frequently used in RE : system-as-is and system-to-be [27]. The former means the system as it exists now while the latter means the system as we want it to be. The role of RE is to identify requirements that will change the system from the as-is state to the to-be state. We consider that a system is a set of components (human, software, hardware…) interacting with each other to satisfy a purpose.

Nuseibeh and Easterbrook [37] proposed a RE process that consists of six activities: elicitation, modelling, analysis, specification, validation and management of requirements. Other authors, in particular Kotonya and Sommerville [34] have highlighted all the above activities except modelling activity which was included in the elicitation activity. The standard (ISO/IEC/IEEE/29148:2011) [38], proposed four activities for the RE process which are: defining the requirements of the stakeholders, requirements analysis, verification and validation, and finally the requirements management.

These differences are not necessarily justified by an omission or addition of activities, but can be considered as different ways of seeing the process. In the following, the activities of the RE process are described based on [27], [36] and [38].

Domain Understanding and Requirements Elicitation. Domain understanding consists of studying the system-as-is within its organizational and technical context. It leads to understand the domain in which the problems are rooted and identify the roots of the problems [27]. As a result:

- Stakeholders involved in the RE process must be identified;
- A comprehensive picture, of the organization's objectives, actors, roles and dependencies among them , in which the system-as-is takes place, is formed;

- The scope of the system-as-is is defined (objectives, components, information flowing through it and constraints);
- Strengths and weaknesses of the system-as-is, as perceived by the identified stakeholders are determined;
- A glossary of terms should be established to provide definitions of key concepts on which everyone should agree.

This result will be utilized for the rest of RE activities. Once the requirements engineer acquires some knowledge about the domain, he starts eliciting requirements. Elicitation is "a cooperative learning process in which the requirement engineer and the system stakeholders work in close collaboration to acquire the right requirements. This activity is obviously critical. If done wrong, it will result in poor requirements and, consequently in poor software" [27]. In this activity, the requirement engineer aims to collect, capture, explore and model the requirements of the system-to-be from a multitude of sources. Modelling is important in this activity because the system needs to be represented faithfully, so that this representation can be understandable by users.

To perform the elicitation activity, a variety of techniques exists: interviews, questionnaires, surveys, prototyping, observation… These techniques has been classified in [27], [37].

Evaluation and Agreement. This activity aims to examine and interpret the elicitation phase results, in order to:

- Clarify the requirements, remove inconsistencies and ensure completeness and non-redundancy;
- Identify and resolve conflicting concerns;
- Assess and resolve risks associated with the system that is being shaped;
- Compare the alternative options identified during elicitation with regard to quality objectives and risks, and select best options on that basis;
- Prioritize requirements in order to resolve conflicts or avoid exceeding budget and deadlines etc…

To support the evaluation activity, a variety of qualitative and quantitative techniques is presented in [27].

Specification and Documentation. The agreed requirements emerging from the evaluation activity must be detailed, structured and documented in the specification document. So that they can be understood by all users involved in the RE process. Specification can be formal, semi-formal or informal, see [27] for more details. The specification document is the main product of RE [36], [38]. It traces the process and includes descriptions of various elements, techniques and tools that have led to the result. Requirements must be classified by users to prepare the validation step [38].

Requirements Consolidation. Also called validation activity, as referred by [37,38]. Requirements engineer detects and corrects errors. He certifies that the requirements meet the expectations of users, and define the expected functionality of the system. A variety of verification method is proposed by the standard (ISO/IEC/IEEE 29148: 2011) [38]. Among the products of this step: a corrected version of the requirements produced by the previous activity; a set of acceptance test sets produced from the requirements specification; and an eventual prototype of the system-to-be.

Requirements Evolution. This activity considers the different versions of requirements. Indeed, Requirements may change due to different causes. Thus requirements before change and after change as well as the causes of change have to be noted in the specification document. Therefore a new version of this document is produced at each change. In [30], [38], a whole process for change management is proposed. Requirements change is inevitable, it should be anticipated from the beginning as well as requirements traceability should be maintained. The former is guaranteed by assigning an attribute to each requirement, in order to specify whether it is stable or may change. The latter, has to be planned from the beginning of the project for two reasons: (i) trace the evolution of requirement and justify any change and (ii) track back the requirement into the initial objectives so that one can argue that they are satisfied.

These activities compose the classical RE process, which emanate from the software engineering field. A DW can be seen as a software system having the specificity of supporting decision making. Engineering the requirements of DWs is a step of DW design known as requirement analysis for DW design. In the next section, this step is discussed through goal-oriented approaches. A set of criteria to evaluate these approaches are described, and comparative analysis is made among six famous goal-oriented approaches.

3 RE for DWs

In this section, a link is made between RE process seen above (section 2), and requirement analysis for DW design. First of all, it is clear that the system-as-is, is represented by the organization before building the DW, while the system-to-be is the DW within the organization. Second, talking about the RE dimensions mentioned above “why, what, who and how” (subsection 2.1); the “why” dimension concerns identifying the high-level objectives and goals of the stakeholders and decision makers involved in the DW development project [16]. While the “what” dimension, is concerned by identifying what information is relevant for decision making [18]. We call that “useful information” for decision making, which should be stored in the DW. The “who” dimension, cares about identifying the stakeholders and decision makers involved in the DW development project. Finally the “how” dimension is not introduced in DWs. We assume that it is concerned about implementation constraints to be applied on the DW. The concept requirement introduced above (section 2) represent, for DW, information requirements that supports decision-making [19], [33],

[39]. In the following, we use the term requirement to refer to information requirement.

Despite the large number of goal-oriented DW design approaches proposed, as yet, there is no common strategy of requirement analysis in DW design [23]. Besides, we argue that there is no common RE process for DW. [40] Proposed a set of activities for goal-oriented approaches, with various models for each activity. This work was exploited by [41] in a comparative study of goal-oriented DW design approaches. The authors evaluated these approaches according to the models used in each step of requirements analysis. Our purpose, is to extract the invariant steps from the classical RE process to be applied in DW requirement analysis, in order to identify a set of criteria to evaluate goal-oriented DW design approaches. In the following subsections, those criteria will be described, and will be used to compare the goal-oriented DW design approaches. Then we give a brief description of the compared approaches, and discuss the result.

3.1 Evaluation Criteria

The context of RE for DW is specific, since DW is dedicated to decision making [33]. We assume that the classical RE process is not completely applied in DW requirement analysis. Thus, in order to see what are the current practices in this field, we studied this classical RE process in the context of DW, and extracted a set of evaluation criteria, then assigned for each criterion a coefficient that reflects its weight in the process of requirements analysis for DWs. The assigned weights are of three types:

- Elementary: criterion qualifies an elementary activity of the RE process. (weight 1)
- Important: Elementary and requires great importance. (weight 2)
- Mandatory: Important and must qualify each approach. (weight 3)

The criteria we suggest, include the following:

Elicitation: In goal-oriented DW design approaches, requirement elicitation is the most complex activity [41] for the following reasons: in one hand DWs are used exclusively for decision making [15], [19], [42]. In the other hand, goal-oriented DW design approaches are based on the analysis of high-level goals [27]. The problem, at this level, is in extracting the goals from decision makers. If in case a decision maker knows how to express his goals, which is not often the case, in some other cases, decision makers poorly express their goals, or less, they are not able to formulate them. The requirement elicitation is the first activity of the RE process, the remaining steps depend on it. If the goals are poorly defined, the DW may not meet the needs of decision makers. Considering the importance of this phase it will have weight (2).

Specification: this criterion qualifies the specification activity, where the elicited goals are analyzed (conflict detection, errors, redundancy) and modelled. The concerns of the requirements engineer is to find, according to the decision makers, which models may be used to specify their needs so that they can be able to understand. It is about mapping the real-world needs into a requirements model [40]. It is a core activity for the RE process and prepares for validation step, therefore it bear the weight (2).

Validation: A consensus on the elicited goals, between the requirement engineer and decision makers must be established through validation. Validation of requirements is paramount for further stages of DW design. If requirements are not validated by decision makers, the risk that the DW will not address their needs increases, which will bring the project to failure. Therefore, validation is mandatory and deserve the weight (3).

Requirements' evolution management: One will not flee the fact that requirements evolve throughout the requirement analysis in DW design. Besides, it's not impossible that they evolve even after validation. A DW not taking into account the evolving requirements is certainly not at the same effectiveness as another one supporting it. Furthermore, a decision maker, always, seeks to meet its objectives in one way or another. Elsewhere, it does not include the fact that he succeeds to express all his needs. It is important to plan, from the beginning of requirement analysis, for alternatives to the defined requirements [27], [42]. Also, anticipate requirements subject to change or evolution. This criterion represent the requirement evolution activity in the classical RE process. Besides, regarding its contribution to the effectiveness of DW, this criterion deserves the weight (2).

Traceability: How will it be possible to affirm that a goal is satisfied? How to define to which goal is associated a given requirement? To answer this, traceability is introduced. It consists on tracing the path from the goal to the relevant information in DW [43]. Traceability helps assessing the impact of changes and rationale comprehension, by identifying which parts of the implementation belong to which requirement [44]. It also supports the reusability and maintainability of DW, since the scope of each part of the project is known and defined thanks to the traces. In turn, these benefits help lowering the costs associated with the project [45,46]. Distinction is made between post-traceability and pre-traceability [47]. The former is about the traceability of the requirement, its deployment, and its use. Whereas the latter is the traceability of a requirement back to its origin which is goal in our context. Thus, since the first RE's task, it is essential to think about keeping trace of everything. This is necessary to justify delays and possibly identify the cause of failures. For all these reasons, the weight (3) is the most suitable for this criterion.

Reusability: DW implementation is a complex and costly activity in resources and time [48]. It also requires specific developments to the characteristics and needs of the organization. However, decision-making projects for the same field of activity or even different business areas have similarities [49]. It is certainly possible to find situations which we have already faced; avoid falling in unrealistic requirements on the basis of earlier experiences; or even propose to decision makers new requirements through anticipation [8]. Reusing requirements, or reusing existing Data marts [8] or even DWs, promote saving time and reliability in future projects. Therefore, this is elementary for each approach and carries the weight (1).

3.2 Comparative Analysis

Six famous goal-oriented DW design approaches have been studied. The study consists of capturing the satisfied criteria for each approach. In the following, these approaches are briefly described.

1. (Bonifati & al 2001) [15]: the approach starts by gathering information from business analysts and/or managers about the company goals and needs [15]. This is accomplished through the Goal/Question/ Metrics paradigm. The goals obtained are aggregated and refined, until a small number of goals subsume all the collected information. Each aggregated goal is specified by an abstraction sheet, which expresses its characteristics in great detail. From abstraction sheets, it is possible to extract the specifications of ideal star schemas which represent users' information requirements for the DW.
2. (Paim & al 2003) [42]: The approach is named DWARF. The authors adopted the classical RE process for DW, adding traceability and compliance of requirements. DWARF is divided into a series of well-defined stages. Each stage presented in a development cycle, applies different levels of abstraction that detail the application more deeply each time, with the goal of creating a baseline for requirements. The latter are specified for data marts and grouped to specify the DW .The authors insisted on the documentation activity of each step of the approach.
3. (Gam & Salinesi 2006) [8]: the approach is called CADWA. In CADWA, the information requirement are extracted from (i) the goals presented by the strategic plan of the organization, (ii) decision makers business plans, (iii) transactional systems, and (iv) the existing DW or data marts models that can be reused. In the next stage the authors create a model using the MAP goal model, to represent the current and future information requirements of decision makers.
4. (Mazon & al 2007) [33]: This approach starts by identifying a hierarchy of goals with tree levels of abstraction: Strategic goals which are fulfilled by decisional goals which are in turn fulfilled by information goals. From information goals, the information required for decision making is directly derived. The authors adapted the strategic rational (SR) model of the i* modelling framework [50] and used it to specify goals and information requirements of decision makers. Traceability does not appear explicitly in the approach. However, the models used allowed pre-traceability of requirement from goals to information requirements.
5. (Giorgini & al 2007) [17]: the approach is called GRanD. It adopts two perspectives for DW requirement analysis. The former is organizational modeling centered on stakeholders and aims to shape the organization. The latter is decisional modeling which is directly related to the information needs of decision makers [17]. Traceability is not made explicit by the approach. However, GRanD is based on an adapted i* modelling framework [50]. Therefore, pre-traceability of requirements can be guaranteed through the proposed models.
6. (Prakash & Gosain 2008) [19]: The authors focus first on the context of the organizational goals. Thus, A GDI (Goal Decision Information) model was proposed with three levels of abstraction. It starts by identifying organizational goals. A goal enables to identify the set of decisions that are relevant. For each

decision, a set of required information to make it, is determined. This organizational view is translated into a technical view by the use of the informational scenario. The latter is written for each decision available in the GDI organization scheme, to capture the required information for decision making.

Table 1. A comparative study of the goal-oriented DW design approaches

The approach	Elicitation (2)	Specification (2)	Validation (3)	evolution management (2)	Traceability (3)	Reusability (1)	weight
Bonifati & al 2001	X	X					4/13
Paim & Castro 2003	X	X	X	X	X		12/13
Gam & Salinesi 2006	X	X				X	5/13
Mazon & al 2007	X	X			X		7/13
Giorgini & al 2007	X	X			X		7/13
Prakash & Gosain 2008	X	X			X		7/13

A set of conclusions is made on basis of table 1. First, all the approaches focus on the elicitation and specification activities of the RE process. These two activities are basic for the RE process. Second, validation criterion which represents validation activity, has not shown great importance from the approaches. It is mentioned above that it is of great importance (section 3.1). Besides, it refers to a basic activity of the classical RE process. Consequently this criterion needs more importance for next approaches. Third, traceability is not well addressed. It is made implicit by the models proposed. More efforts has to be made to satisfy that criterion, due to its contribution to the proper conduct of the RE process. Forth, requirement evolution management

criterion is only satisfied by the DWARF approach. It was addressed by a horizontal activity since the beginning of the approach until the end. Finally, concerning reusability criterion, only CADWA [8] applied it by reusing existing structures of DWs or data marts.

DWARF [42] has encompassed the large number of criteria since it applied the classical RE process. Consequently, it has the highest weight among the approaches. [17], [19] and [33] has well addressed the elicitation and specification criteria. This what made of them powerful approaches, but still, they have to incorporate validation activity in the process of the approach, and plan for a better traceability.

4 Conclusion

In this paper, a comparative study was made among goal-oriented DW design approaches. We have investigated to what extent these approaches went well with respect to the classical RE process. Our study was based on six evaluation criteria, which were defined directly from the RE process for many reasons. We argue that a DW is more than a software system, it has the specificity of providing useful information to support decision-making. Thus, RE process for DWs has to be applied carefully. In addition, there is no standard approaches for DW design despite the considerable efforts made in the field. The main motivation of this work is to serve as a starting point for researchers to think at developing a standard RE process for DW design. Consequently, this comparative study can be useful for researchers in achieving a common understanding in the field and providing a solid foundation for the research community.

References

1. Inmon, W.H.: Building the data warehouse, 2nd edn. (1996)
2. Golfarelli, M., Maio, D., Rizzi, S.: The dimensional fact model: a conceptual model for data warehouses. International Journal of Cooperative Information Systems 7, 215–247 (1998)
3. Hüsemann, B., Lechtenbörger, J., Vossen, G.: Conceptual data warehouse design. In: Proceedings DMDW, Stockholm, Sweden, pp. 3–9 (2000)
4. Moody, D., Kortink, M.: From enterprise models to dimensionalmodels: a methodology for data warehouse and data mart design. In: Proceedings DMDW, Stockholm, Sweden (2000)
5. List, B., Bruckner, R.M., Machaczek, K., Schiefer, J.: A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse. In: Hameurlain, A., Cicchetti, R., Traunmüller, R. (eds.) DEXA 2002. LNCS, vol. 2453, pp. 203–215. Springer, Heidelberg (2002)
6. Bruckner, R., List, B., Scheifer, J.: Developing Requirements for Data Warehouse Systems with Use Cases. In: AMCIS 2001 Proceedings, pp. 329–335 (2001)
7. Winter, R., Strauch, B.: A Method for Demand-driven Information Requirements Analysis in Data Warehousing Projects. In: The Hawai'i International Conference on Systems Sciences, January 6-9 (2003)

8. Gam, I., Salinesi, C.: A Requirement-driven Approach for Designing Data Warehouses. In: Foundations for Software Quality, Luxembourg, pp. 1–15 (June 2006)
9. Prakash, N., Gosain, A.: Requirements Driven Data Warehouse Development. In: 1st JIIT A10, Sector 62 NOIDA 201307, India (2003)
10. Golfarelli, M.: Data warehouse life-cycle and design. In: Encyclopedia of Database Systems, pp. 658–664. Springer (2009)
11. List, B., Schiefer, J., Tjoa, A.M.: Process-Oriented Requirement Analysis Supporting the Data Warehouse Design Process A Use Case Driven Approach (2000)
12. Schiefer, J., List, B., Bruckner, R.: A holistic approach for managing requirements of data warehouse systems. In: AMCIS 2002 Proceedings, vol. 13 (2002)
13. Kimball, R., Ross, M.: The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons (2011)
14. 24765:2010, ISO\IEC\IEEE: Systems and software engineering Vocabulary (2010)
15. Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S.: Designing Data Marts for Data Warehouses. ACM Transactions on Software Engineering and Methodology 10(4), 452–483 (2001)
16. Giorgini, P., Rizzi, S., Garzetti, M.: Goal-Oriented Requirement Analysis for Data Warehouse Design (2005)
17. Giorgini, P., Rizzi, S., Garzetti, M.: GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decision Support Systems 45, 4–21 (2007)
18. Mazón, J.-N., Trujillo, J.: An MDA approach for the development of data warehouses. Decision Support Systems 45, 41–58 (2008)
19. Prakash, N., Gosain, A.: An approach to engineering the requirements of data warehouses. Requirements Eng. 13, 49–72 (2008)
20. Rolland, C.: Reasoning with goals to engineer requirements. In: Enterprise Information Systems V, pp. 12–20. Springer (2005)
21. Giorgini, P., Mylopoulos, J., Nicchiarelli, E., Sebastiani, R.: Formal reasoning techniques for goal models. Journal on Data Semantics I, 1–20 (2003)
22. Stefanov, V., List, B.: Business Metadata for the DataWarehouse. In: 10th IEEE International Enterprise Distributed Object Computing Conference Workshops, 2006, p. 20. IEEE (2006)
23. Cravero Leal, A., Mazón, J.N., Trujillo, J.: A business-oriented approach to data warehouse development. Ingeniería e Investigación 33, 59–65 (2013)
24. Lamsweerde, A.V.: Requirements engineering in the year 2000: A research perspective. In: 22nd International Conference on Software Engineering, Invited Paper. ACM Press (2000)
25. Mylopoulos, J., Borgida, A., Yu, E.S.K.: Representing Software Engineering Knowledge. Automated Software Engineering 4, 291–317 (1997)
26. Ross, D.T., Schoman, K.E.: Structured Analysis for Requirements Definition. IEEE Transactions on Software Engineering 3, 10 (1977)
27. Van Lamsweerde, A.: Requirements engineering: from system goals to UML models to software specifications (2009)
28. Rolland, C., Prakash, N.: From Conceptual Modelling to Requirements Engineering. Annals of Software Engineering on Comparative Studies of Engineering Approaches for Software Engineering (2001)
29. Zave, P., Jackson, M.: Classification of Research Efforts in Requirements Engineering. ACM Computing Surveys 29, 7 (1997)
30. Dardenne, A., Lamsweerde, A.V., Fickas, S.: Goal-directed requirements acquisition. Science of Computer Programming 20, 3–50 (1993)

31. Rolland, C., Grosz, G., Kla, R.: Experience With Goal-Scenario Coupling in Requirements Engineering. In: Fourth IEEE International Symposium on Requirements Engineering (RE 1999), June 7-11 (1999)
32. Si-Saïd, S., Rolland, C.: Formalising guidance for the crews goal-scenario approach to requirements engineering. In: The 8th European - Japanese Conference on Information Modelling and Knowledge Bases, May 25-29, p. 20 (1998)
33. Mazón, J.-N., Pardillo, J., Trujillo, J.: A model-driven goal-oriented requirement engineering approach for data warehouses. In: Hainaut, J.-L., et al. (eds.) ER Workshops 2007. LNCS, vol. 4802, pp. 255–264. Springer, Heidelberg (2007)
34. Kotonya, G., Sommerville, I.: Requirements Engineering: Process and Techniques (1998)
35. Loucopoulos, P., Karakostas, V.: System Requirements Engineering. McGraw-Hill Book Company Europe (1995)
36. Rolland, C.: De la modélisation conceptuelle à l'ingénierie des exigences. Journal Techniques de l'Ingénieur, 23 (2011)
37. Nuseibeh, B., Easterbrook, S.: Requirements Engineering: A Roadmap. In: International Conference on Software Engineering, pp. 4–11. ACM Press (juin 2000)
38. 29148:2011, ISO\IEC\IEEE: Systems and software engineering — Life cycle processes — Requirements engineering, pp. 95 (2011)
39. Winter, R., Strauch, B.: Information Requirements Engineering for Data Warehouse Systems. In: ACM Symposium on Applied Computing, SAC 2004, Nicosia, Cyprus, March 14-17, pp. 1359–1365 (2004)
40. Kavakli, E., Loucopoulos, P.: Goal driven requirements engineering: evaluation of current methods. In: Proceedings of the 8th CAiSE/IFIP8, pp. 16–17 (2003)
41. Leal, A.C., Sepúlveda, S., Mate, A., Mazón, J.-N., Trujillo, J.: Goal oriented requirements engineering in data warehouses: A comparative study. Ingeniería E Investigación 34, 66–70 (2014)
42. Paim, F.R.S., de Castro, J.F.B.: DWARF: An approch for requirement definition and management of DW systems. In: 11th IEEE International Requirement Engeneering Confrence (2003)
43. Hull, E., Jackson, K., Dick, J.: Requirements Engineering, 2nd edn., p. 201. Springer (2005)
44. Antoniol, G., Canfora, G., Casazza, G., De Lucia, A., Merlo, E.: Recovering traceability links between code and documentation. IEEE Transactions on Software Engineering 28, 970–983 (2002)
45. Ramesh, B., Stubbs, C., Powers, T., Edwards, M.: Requirements traceability: Theory and practice. Annals of Software Engineering 3, 397–415 (1997)
46. Ramesh, B., Jarke, M.: Toward reference models for requirements traceability. IEEE Transactions on Software Engineering 27, 58–93 (2001)
47. Pohl, K.: PRO-ART: Enabling requirements pre-traceability. In: Proceedings of the Second International Conference on Requirements Engineering, pp. 76–84 (1996)
48. Carneiro, L., Brayner, A.: X-META: A Methodology for Data Warehouse Design with Metadata Management. In: Design and Management of Data Warehouses (DMDW), pp. 13–22 (2002)
49. Annoni, E., Ravat, F., Teste, O., Zurfluh, G.: Les systèmes d'informations décisionnels : une approche d'analyse et de conception à base de patrons. Revue des Sciences et Technologies de l'Information, série ISI, «Méthodes Avancées de Développement des SI » 10, 81–106 (2005)
50. Yu, E.S.: Towards modelling and reasoning support for early-phase requirements engineering. In: Proceedings of the Third IEEE International Symposium on Requirements Engineering, pp. 226–235. IEEE (1997)

Information Technology: OLAP and Web Services

Research and Analysis of the Stream Materialized Aggregate List

Marcin Gorawski(✉) and Krzysztof Pasterak

Silesian University of Technology,
Institute of Computer Science,
Akademicka 16, 44-100 Gliwice Poland
{Marcin.Gorawski,Krzysztof.Pasterak}@polsl.pl

Abstract. The problem of low-latency processing of large amounts of data acquired in continuously changing environment has led to the genesis of Stream Processing Systems (SPS). However, sometimes it is crucial to process both historical (archived) and current data, in order to obtain full knowledge about various phenomena. This is achieved in a Stream Data Warehouse (StrDW), where analytical operations on both historical and current data streams are performed. In this paper we focus on Stream Materialized Aggregate List (StrMAL) – a stream repository tier of StrDW. As a motivating example, the liquefied petrol storage and distribution system, containing continuous telemetric data acquisition, transmission and storage, will be presented as possible application for Stream Materialized Aggregate List.

Keywords: Materialized aggregate list · Stream data warehouse · Stream processing

1 Introduction

Nowadays, the necessity of processing, storing and analyzing of very large data volumes (considered also as *BigData*) is constantly growing. This implies development of newer and more advanced systems, that are able to satisfy this need. Moreover, from the perspective of various enterprises, organizations and other data producers and consumers, the outcome information is expected to be reliable, most up-to-date and obtained in the shortest time possible. These requirements determine the attractiveness of solutions already present on market, as well as constitute new objectives for developers [22].

In the following paper we focus on Stream Data Processing Systems (SPS). They are designed to process current and continuously generated data with relatively high frequency. When non-stream solutions (i.e. those relying on persistent data) are concerned, processing unit enforces collecting data from sources. Stream oriented systems have to process incoming data almost instantly as they arrive, since data are produced and actively delivered by sources. There are representative examples of Stream Processing Systems [1–6, 26, 27], however they are relatively not as popular as classic, traditional data storage systems.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 269–278, 2015.
DOI: 10.1007/978-3-319-19578-0_22

The example of application involving instant and immediate analysis of data delivered continuously is a liquefied petrol storage and distribution system. Such an installation consists of multiple petrol stations, where various measurements are gathered and transmitted to the centralized or distributed analysis platform.

Each petrol station is equipped with fuel tanks where liquefied fuel is stored and dispensing devices which act as sale endpoints. These appliances generate two streams of data supplemented with delivery records entered by station workers or detected automatically. Usually fuel volume and temperature is measured in tanks, whereas the amount of sold fuel is returned from meters installed in dispensers.

The common analysis performed upon the aforementioned values aims to detect various anomalies and other adverse phenomena that can occur at petrol stations. The most dangerous example is fuel leak [15, 24], which introduces very serious consequences to the environment. In order to prevent such a threat, it is crucial to detect any volume of fuel leaked from tank and piping as fast as possible.

This paper is organized as follows. Section 2 contains information concerning data stream storage problems with theoretical base of a Stream Materialized Aggregate List (StrMAL) described. In Section 3 the architecture of StrMAL is presented along with examples of its most important features. Section 4 contains test results performed over a StrMAL engine, whereas Section 5 summarizes the paper.

2 Data Stream Storage

A Data stream [8, 12–14] can be defined as an infinite sequence of tuples with unique timestamps and attributes carrying information describing various phenomena at subsequent moments of time. Stream Processing Systems usually do not provide any storage operation in their work flow, since they are designed to produce answers immediately as new data arrive. Optional data storage is sometimes used to provide static data as an extension to stream data.

Under certain circumstances an instant access to historical data stored in a database, as well as efficient processing performed on current data is required. Analyzing the history is necessary in learning process, where different trends, dependencies, and rules are discovered and remembered [21]. Later, gathered knowledge is used to filter current stream in order to detect any desired events. This process frequently involves browsing data on a certain level of detail – in other words – on different aggregation levels. Moreover, data retrieval and aggregation operations should not interfere with insertion of newly arrived data, which often cannot be completely eliminated.

The problem of processing both stored and current data has lead to the idea of the Stream Data Warehouse [7, 9, 19, 20, 23, 25]. It is an unified processing platform capable to produce immediate answers to complex queries concerning current and archived data. In current mode, data can be processed before they are persisted in any data structure, as in Stream Processing Systems.

2.1 Problems and Issues

As a consequence of data stream nature, it is virtually not possible to store a whole stream in a memory. In addition, at a given moment of time, the stream contains only the most current tuples, since all read before have been already removed and archived, which forces searching the history (database).

Moreover, data in a stream are produced relatively frequently and in a unpredictable manner, which causes database to be updated very often and irregularly. High intensity of modifying transactions, being made in parallel with queries consisting of large range data retrieval, can lead to serious decrease in overall performance.

Relative database systems are designed to execute versatile CRUD (Create, Read, Update and Delete) operations on the whole dataset stored inside their internal memory. However, in stream appliances, updates take place only at the end of a time frame, i.e. tuples arrive and are organized ascending by timestamps. In this cause, it is not possible that once stored piece of data is updated.

Repetitive execution of the same or similar operations (e.g. aggregation) of the same datasets is usually time consuming and thus leading to unnecessary delays. In order to prevent these adverse situations, results of time costing operations can be stored along with query parameters to provide access to once computed values. As mentioned before, there are no updates on historical data causing the materialized data to be immutable.

2.2 Stream Materialized Aggregate List

Many items that are sequentially arranged (as tuples in a data stream) can be stored in a list data structure. In such a form, it is easy to view all subsequent elements in proper order. When browsing tuples from a stream, consecutive retrieving is the only operation considered here, which can be described as forward iterating over list.

An aggregate list [16–18] can be defined as a sequential data structure, containing a subset of an aggregated stream (stream of aggregated tuples). It is stored in memory and acts as a physical representation of stream, beginning from certain moment of time. Because of the limited capacity of list, it is assumed that all aggregates already read can be replaced with more fresh data.

When considering various data collections, extracting data access operations into a separate interface is a common practice. Such an interface is called an *iterator* and is used to traverse any data structure (as aggregate list for example). Thus, an iterator can be used for retrieving tuples from a stream.

The aforementioned issues became a motivation for designing a solution which is capable to provide an uniform access to any data stream, efficiently manage available memory, and avoid redundant operations. It is done by using aggregate list, iterator interface, and aggregate materialization techniques. The solution has been named a Stream Materialized Aggregate List [11] and is designed to act as a data storage tier of a Stream Data Warehouse [10].

It is possible to create several iterators attached to a single list and pointing to different elements. In such a situation, the distance (measured in time units)

between them is unconstrained and can be arbitrary long, causing the whole list to occupy very large amount of memory. In order to prevent this situation, the following solution is used: the aggregate list itself is not located in memory, instead its active fragments (tuples being currently in use) are stored inside each iterator.

Moreover, in order to increase memory management efficiency, the following solution is used [17, 18]: each iterator contains a static array which corresponds to an iterator-specific aggregate list fragment. As far as successive aggregates are retrieved from an iterator and become outdated, they are replaced by newer ones. Each array is logically divided into pages (basic units). Due to that, certain number of ready-to-read aggregates is always available. When the need of new aggregates creation occurs, a whole page at once is produced and replaces the old one.

3 Architecture of StrMAL

The Stream Materialized Aggregate List was implemented using multilayer concept. Each of them is realizing separate functionalities and is responsible for another stage of aggregates production. Figure 1 presents the overall architecture of StrMAL. Four layers have been denoted by the following acronyms: SDL, APL, DML, and CL – they are described later in the text.

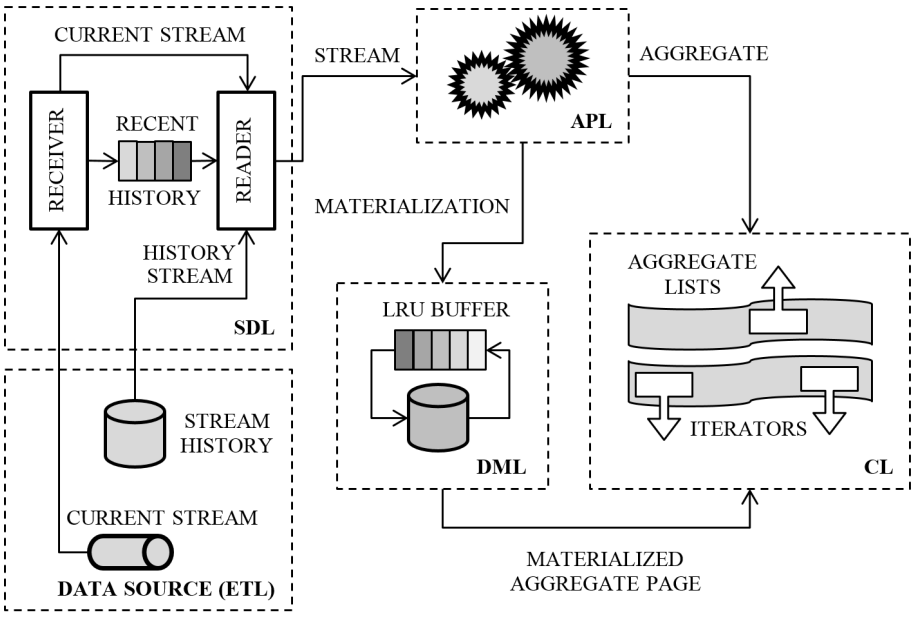


Fig. 1. Architecture of the StrMAL engine

In the Stream Distribution Layer (SDL) the process of aggregate list production begins with collecting data required for aggregation. It is done, depending on start time specified in query, by using current or historical stream. This layer provides a uniform access to data streams, irrespective of their origin (source) and start time.

The Aggregate Production Layer (APL) retrieves desired streams from the SDL and, basing on parameters obtained from client, performs aggregation. Outcome aggregates are delivered to clients and materialized (persisted for future use).

The Data Materialization Layer (DML) involves persisting aggregates in database, along with query parameters. Besides storing, this layer also provides searching and retrieving mechanisms. Cache memory (LRU buffer) is used to achieve better performance of I/O operations.

The Client Layer (CL) is responsible for communication with clients and providing them aggregates produced in the APL with data retrieved from the SDL or materialized aggregates read from the DML. It integrates all mentioned layers and uses them to prepare, produce, and serve results.

3.1 Current and Historical Stream Support

One of the major tasks of the Stream Distribution Layer is to collect tuples from current stream and store them temporally in a buffer called History Table (HT). It is performed in order to provide a flexible bridge between current and historical data. When the SDL is queried for a data stream beginning from a certain timestamp, first it performs a lookup over the HT to determine whether the desired data have been already produced – and when it is true – if they are still in the HT or have been persisted in a database.

The Stream Distribution Layer can operate in four states, depending on the distance between current and searched time. Each state determines the source from which data are retrieved and other working principles, such as next state reached under state-specific circumstances. These states are named as follows:

1. TAB – tuples are read from the History Table,
2. DB – tuples are read from a database,
3. SYNC – synchronization with the current stream,
4. CUR – tuples are read from the current stream.

The TAB is the starting state, when the SDL is queried with a specific timestamp and the History Table is searched to find whether the desired tuple is present in it. When the tuple is not found in the HT, it either has not been generated and acquired by the system or it has been archived and stored in a database. In the former case the SDL remains in the TAB state waiting for tuple to appear in the HT, whereas in latter cause, the SDL switches into the DB state and reads tuples from a database until the end of batch (certain number of tuples) is reached. After that, the SDL switches back into the TAB state.

In the other situation, after successful lookup, SDL remains in TAB state until the end of HT is reached (there are no more tuples to read). Such a circumstance denotes that next tuple ought to be retrieved from the current stream. However this process cannot simply be performed by switching into CUR state (when subsequent tuples are read from the current stream). SDL needs to synchronize with the current stream – it is done by entering SYNC state. In that state the SDL assures that no tuples will be omitted during switching – i.e. tuples being removed from the current stream and not yet written into the HT. Figure 2 presents state diagram of the Stream Distribution Layer.

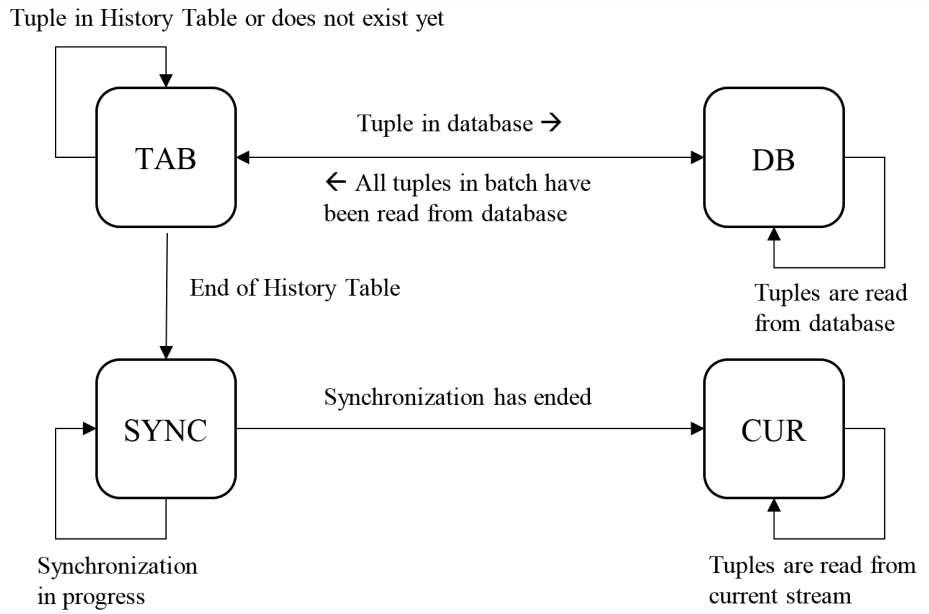


Fig. 2. State diagram of the SDL

4 Test Results

First test was conducted in order to verify the impact of History Table on archived tuple read time. The objective to that study was to simulate the situation involving reading subsequent tuples from current stream with variable time gap (delay) between each read operation. In such a case client reading tuples desynchronizes with stream and is obliged to perform a lookup in database containing archived data. When HT is used, it stores recent history of stream and allows the client of SDL to retrieve desired data from buffer instead of database.

Figure 3 presents tuple read time (in microseconds) with buffering in HT applied depending on HT size (in number of tuples). Three different delays were used: 1 s, 2 s, and 5 s. The starting size of HT was set to 8 and it was doubled respectively when there were any calls to database. The test was finished when all tuples were read from the buffer allowing client not to operate on persistent storage at all. Results show that tuple read time when using HT in 100% is about 6 times shorter than in the 5 s delay example (where almost 100% of read operations were made on a database).

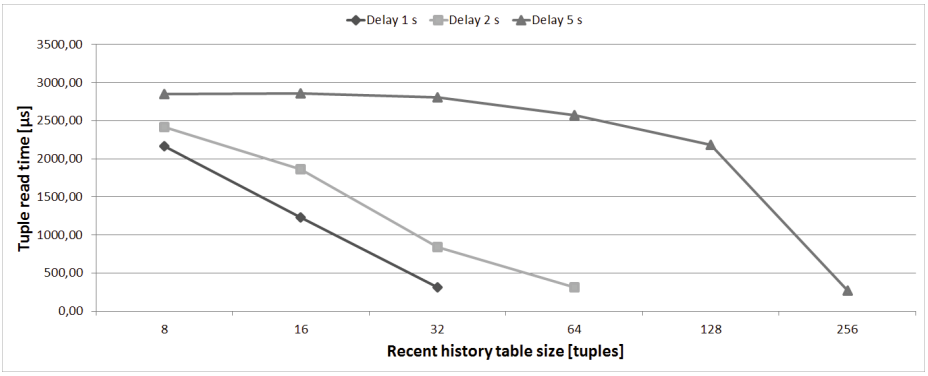


Fig. 3. Tuple read time depending on HT size

Next test was conducted to verify the percentage of HT calls in historical data retrieval depending on aggregate consumption time. Aggregates were read from CL and the following delays were introduced: 300 ms, 400 ms, 500 ms, 550 ms, 600 ms, 750 ms, and 1000 ms. Such values have been selected after preliminary tests which showed that below 300 ms there are no calls to any historical data because every tuple is read directly from the current stream. Between 500 ms and 600 ms an additional measure was performed (at 550 ms) due to high variability in that range. Four different sizes of History Table were used: 8, 16, 64, and 256 (measured in number of tuples).

Figure 4 shows that for two first examples (HT sizes: 8 and 16) the percentage of calls to HT suddenly dropped from 100% to about 10% at delay set to about 600 ms. It means that 90% of calls to historical data sources were made to database causing the overall aggregate production time to be longer. When HT size was set to 256 tuples about 40% of calls were still made to HT, even when aggregate consumption time was equal to 1 s.

Results of performed tests showed that using the History Table as buffering mechanism, while performing seamless switching between historical and current data sources, is legitimate. Tuple read time is noticeably shorter and database system is less loaded causing the whole process of aggregate production more efficient.

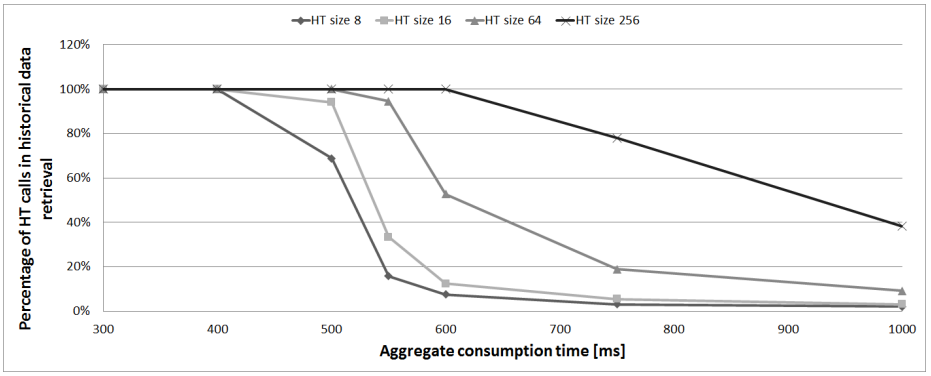


Fig. 4. Percentage of HT calls depending on aggregate consumption time

5 Summary

In this paper we have described the architecture of the Stream Materialized Aggregate List engine, which is a component of Stream Data Warehouse, responsible for storing and serving data streams on various levels of aggregation. The StrDW itself is still at the planning stage, while its components, such as described StrMAL engine, are being intensively developed and tested.

In the nearest future we intend to design all concepts and modules of the StrDW, with spatial indexing, distributed architecture and low-latency query processing issues included. The target system is expected to process data streams in OLAP manner, allowing the analysis on currently changing multidimensional aggregated data to be performed in decision supporting applications with critical time requirements with distributed environment and concurrency issues involved, such as the aforementioned liquefied petrol storage and distribution system.

References

1. Abadi, D.J., Ahmad, Y., Balazinska, M., Çetintemel, U., Cherniack, M., Hwang, J.-H., Lindner, W., Maskey, A., Rasin, A., Ryvkina, E., Tatbul, N., Xing, Y., Zdonik, S.B.: The design of the borealis stream processing engine. In: CIDR, pp. 277–289 (2005)
2. Abadi, D.J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., Zdonik, S.: Aurora: A new model and architecture for data stream management. The VLDB Journal 12(2), 120–139 (2003)
3. Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Datar, M., Ito, K., Motwani, R., Srivastava, U., Widom, J.: Stream: The stanford stream data manager. Technical Report 2003-21, Stanford InfoLab (2003)
4. Arasu, A., Widom, J.: A denotational semantics for continuous queries over streams and relations. SIGMOD Rec. 33(3), 6–11 (2004)

5. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and issues in data stream systems. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2002, pp. 1–16. ACM, New York (2002)
6. Barga, R.S., Goldstein, J., Ali, M.H., Hong, M.: Consistent streaming through time: A vision for event stream processing. In: CIDR, pp. 363–374, http://www.cidrdb.org
7. Bateni, M., Golab, L., Hajiaghayi, M., Karloff, H.: Scheduling to minimize staleness and stretch in real-time data warehouses. Theory of Computing Systems 49(4), 757–780 (2011)
8. Gilbert, A.C., Kotidis, Y., Muthukrishnan, S., Strauss, M.: Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In: Proceedings of the 27th International Conference on Very Large Data Bases, VLDB 2001, pp. 79–88. Morgan Kaufmann Publishers Inc., San Francisco (2001)
9. Golab, L., Johnson, T., Shkapenyuk, V.: Scheduling updates in a real-time stream warehouse. In: IEEE 25th International Conference on Data Engineering, ICDE 2009, pp. 1207–1210 (2009)
10. Gorawski, M.: Advanced data warehouses. Habilitation. Studia Informatica 30(3B), 386 (2009)
11. Gorawski, M.: Time complexity of page filling algorithms in materialized aggregate list (mal) and mal/trigg materialization cost. Control and Cybernetics 38(1), 153–172 (2009)
12. Gorawski, M., Chrószcz, A.: The design of stream database engine in concurrent environment. In: OTM Conferences (2), pp. 1033–1049 (2009)
13. Gorawski, M., Gorawska, A., Pasterak, K.: Evaluation and development perspectives of stream data processing systems. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2013. CCIS, vol. 370, pp. 300–311. Springer, Heidelberg (2013)
14. Gorawski, M., Gorawska, A., Pasterak, K.: A survey of data stream processing tools. In: Information Sciences and Systems, pp. 295–303. Springer International Publishing (2014)
15. Gorawski, M., Gorawska, A., Pasterak, K.: Liquefied petroleum storage and distribution problems and research thesis. In: Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds.) BDAS 2015. CCIS, vol. 521, pp. 540–550. Springer, Heidelberg (2015)
16. Gorawski, M., Malczok, R.: Multi-thread processing of long aggregates lists. In: PPAM, pp. 59–66 (2005)
17. Gorawski, M., Malczok, R.: On efficient storing and processing of long aggregate lists. In: Tjoa, A.M., Trujillo, J. (eds.) DaWaK 2005. LNCS, vol. 3589, pp. 190–199. Springer, Heidelberg (2005)
18. Gorawski, M., Malczok, R.: Towards storing and processing of long aggregates lists in spatial data warehouses. In: XXI Autumn Meeting of Polish Information Processing Society Conference Proceedings, pp. 95–103 (2005)
19. Kakish, K., Kraft, T.A.: Etl evolution for real-time data warehousing. In: 2012 Proceedings of the Conference onInformation Systems Applied Research New Orleans Louisiana (2012)
20. Polyzotis, N., Skiadopoulos, S., Vassiliadis, P., Simitsis, A., Frantzell, N.: Meshing streaming updates with persistent data in an active data warehouse. IEEE Transactions on Knowledge and Data Engineering 20(7), 976–991 (2008)
21. Sigut, M., Alayón, S., Hernández, E.: Applying pattern classification techniques to the early detection of fuel leaks in petrol stations. Journal of Cleaner Production 80, 262–270 (2014)

22. Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. SIGMOD Rec. 34(4), 42–47 (2005)
23. Thiele, M., Bader, A., Lehner, W.: Multi-objective scheduling for real-time data warehouses. Computer Science - Research and Development 24(3), 137–151 (2009)
24. United States Environmental Protection Agency. Preventing Leaks and Spills at Service Stations. A Guide for Facilities (2003), http://www.epa.gov/region9/waste/ust/pdf/servicebooklet.pdf
25. Vassiliadis, P., Simitsis, A.: Near real time etl. In: New Trends in Data Warehousing and Data Analysis. Springer US (2009)
26. Wu, E., Diao, Y., Rizvi, S.: High-performance complex event processing over streams. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD 2006, pp. 407–418. ACM, New York (2006)
27. Zdonik, S.B., Stonebraker, M., Cherniack, M., Çetintemel, U., Balazinska, M., Balakrishnan, H.: The Aurora and Medusa projects. IEEE Data Eng. Bull. 26(1), 3–10 (2003)

SOLAP On-the-Fly Generalization Approach Based on Spatial Hierarchical Structures

Tahar Ziouel(✉), Khalissa Amieur-Derbal(✉), and Kamel Boukhalfa(✉)

High School of Computer Sciences, Algiers
t_ziouel@esi.dz
USTHB University, Algiers
{kderbal,kboukhalfa}@usthb.dz

Abstract. On-the-fly generalization, denotes the use of automated generalization techniques in real-time. This process creates a temporary, generalized dataset exclusively for visualization, not for storage or other purposes. This makes the process well suited to highly interactive applications such as online mapping, mobile mapping and SOLAP. BLG tree is a spatial hierarchical structure widely used in cartographic map generalization and particularly in the context of web mapping. However, this structure is insufficient in the context of SOLAP applications, because it is mainly dedicated to the geographic information processing (geometric features), while SOLAP applications manage a very important decision information that is the measure. In this paper, we propose a new structure, SOLAP BLG Tree, adapted to the generalizaion process in the SOLAP context. Our generalization approach is based on this structure and uses the simplification operator. Combining the topological aspect of geographical objects and the decisional aspect (the measure).

Our experiments were performed on a set of vector data related to the phenomenon of road risk.

Keywords: On-the-fly map generalization · Hierarchical spatial structures · Spatial data warehouses · SOLAP

1 Introduction

Business intelligence is a major decision-making tool for strategic and daily management of data in the enterprise. It provides essential information in several forms to users (decision makers) so that they can analyze and manage their business by taking effectives decisions. Data warehousing and On Line Analytical Processing (OLAP) are technologies intended to support business intelligence. Indeed, Analysts and decision makers in the enterprise can thus analyze interactively and iteratively multidimensional data at a detailed or aggregated level of granularity through online Analytical Processing tools, OLAP, [1] [2] [3]. Nevertheless, these data may have a geographic component that OLAP systems cannot process due their lack of tools for managing spatial data. A new technology has so, emerged, Spatial OLAP (SOLAP), resulting from integrating GIS technology (Geographic Information System) and OLAP [4] [5].

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 279–290, 2015.
DOI: 10.1007/978-3-319-19578-0_23

SOLAP has been defined by [6] *as a visual platform built especially to support rapid and easy spatio-temporal analysis and exploration of data following a multidimensional approach comprised of aggregation levels available in cartographic displays as well as in tabular and diagram displays.*

SOLAP enriches the analysis of classical OLAP systems capabilities in many ways. For instance, providing visual information through maps and interacting with them by formulating queries directly on the cartographic display. Thus, the cartographic component in OLAP systems represents a graphic interface to spatial data warehouses (SDW) which introduce spatial data as subject or analysis axes.

In this context, the analysis of multi-dimensional spatial data often requires navigating through different levels of detail, in order to study the evolution of a phenomenon (fact), and thus allows an effective decision-making. On-the-fly generalization process is therefore well suited to this context, because it can interactively adapt the visualized geographic information to decision-makers needs [7]. However it only addresses the cartographic aspect, at the expense of the decisional one, which is important in multidimensional analysis.

Widely addressed in cartography, on-the-fly generalization, well suited to highly interactive applications such as SOLAP, consists in generating temporary data at different levels of detail from the most detailed level. Different on-the-fly generalization approaches have been developed [8] and classified in two main groups. The first group relies on fast map generalization algorithms that generate coarser levels-of-detail in real-time. [9]. The second group utilizes hierarchical spatial data structures [10] [11]. To the best of our knowledge, no work has proposed a generalization approach for SOLAP, nor the consideration of the decision making aspect (measure) in the generalization process.

In this paper, we propose an on-the-fly generalization approach for SOLAP systems. The approach we propose integrates topographic appearance (distance) and decision-making aspects (measure) for an on-the-fly generalization suited to cartographic experts and decision makers.

The present paper is organized as follows; the next section introduces some research work related to the addressed issue. Section 3 presents a detailed description of the proposed approach. The different steps of our experiments and some results are described in section 4. Section 5 concludes the paper and presents some perspectives.

2 Related Work

Several research work have addressed the generalization for more than three decades [12] [13] [14]. On-the-fly generalization has emerged with the development of highly interactive applications of cartography such as web mapping. The main used operators are selection and simplification [8]. Among research work that have adressed on-the-fly generalization, we can cite [15] [16] carried in the contexte of the European project GiMoDig [17]. The objective of this project is to develop and test methods for providing spatial data to mobile users through

real-time generalization. The work presented in [18], combines multiple representation and cartographic generalization and uses an implementation of multi-agent system where each agent was equipped with a genetic patrimony.

Since cartographic generalization creates a hierarchy of levels of detail, it is natural to use hierarchical structures such as tree structures for storage of the geometry (point, line, polygon) of an object in the highest level of detail. This structure is enriched with information that reflect the importance of a hierarchical level, from which, requested levels of detail may be generated. The generalization process is therefore, speeded up with rapid access to the elaborate structures. For each type of spatial data, corresponds an appropriate hierarchical structure that enables interactive and rapid generalization of geographic objects. BLG tree (Binary Line Generalization tree) has been proposed for linear objects [19] [20], it applies the simplication generalization operator that uses a variant of the Douglas-Peucker algorithm [21]; instead of deleting the less important vertices, it stores them in the structure. The GAP tree (Generalized Area partitionning) has been proposed for the selection and fusion of polygons [20] [10] [11]. The Quadtrees have been proposed for point objects point objects, they allow applying the selection, simplification, aggregation and displacement operators [8] [22].

Furthermore, as we have already mentioned in Section 1, and to the best of our knowledge, there are no research work that have addressed integration of generalization in SOLAP. Nevertheless, some work focused on integration of spatial data as dimension or fact in SDW [4] [7].

In this paper, we propose to integrate on-the-fly generalization process in SOLAP, to adapt the level of detail that meets the decision-makers needs. The approach we propose focuses on linear objects that represent rivers, roads, etc. These latters constitute a geographical dimension linked to the phenomenon of road risk that we consider as use case study. BLG tree structure is dedicated to cartographic generalization of linear objects (roads in our case study). However, this structure cannot be efficiently used in decision-making process, because they don't consider the main decisional information in SOLAP, that is *the measure*.

To better understand this problem, we propose the example illustrated in Figure 1. The analyzed map contains six objects with associated measures. As presented in this example, among the objects at the most detailed level , the object C possesses the greatest measure (30), despite its geometric size is not indicative (see figure 1.a). When reducing the scale, the classical generalization process is triggered, considering only the topographic aspect, the object C is imperceptible (see figure 1.b) despite its decision relevance (the greatest measure) compared to the objective of the analysis performed by the decision-makers.

We propose a generalization approach based on a new version of BLG tree adapted to SOLAP called SOLAP BLG tree.

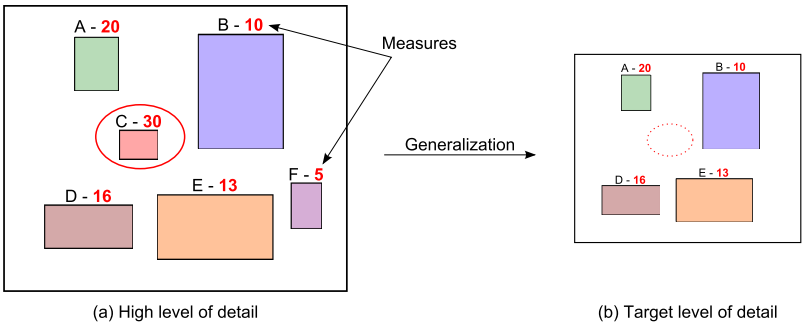


Fig. 1. Traditional generalization results

3 Proposed Approach

The main objective of our approach is to develop an on-the-fly generalization system adapted to SOLAP applications. This system must be able to combine the decision and cartographic aspects to produce maps adapted to the needs of decision makers. Figure 2 shows the overall architecture of our approach. The spatial data warehouse stores decision data (measures, fact, dimensions, etc.) and cartographic data. The latter represented in a single level of detail (the highest one). When the user sends his request, the result is extracted from the stored data. It does not necessarily reflect the level of detail requested by the decision maker, therefore an on-the-fly generalization process is necessary to adapt this result to the expressed need.

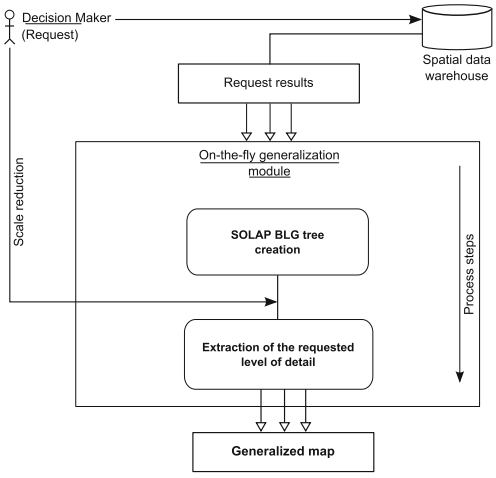


Fig. 2. Global architecture of our approach

The generalization process that we propose is based on SOLAP BLG tree structure. A set of parameters related to the decision aspect is integrated to the structures including the importance function, aggregation etc. All these concepts are described as in the following sections.

3.1 SOLAP BLG Tree

The creation of the SOLAP BLG tree revolves around two main steps: (1) the attribution of an importance value to each point of the polyline and (2) Creating the Hierarchy considering the importance of the points. Indeed, a polyline object (road, river, etc.) consists of a set of points (vertices). SOLAP BLG tree stores these points in a hierarchical structure. Each node of the structure consists of a point of the polyline along with its importance value elaborated by the following function: the importance $I(p)$ for each point p will be determined according to its distance $D(p)$ and its associated measure $M(p)$ as follows: $I(p) = f(D(p), M(p))$. This function can be described by the sum of its cartographic importance (distance) and its decisional importance (measure):

$$I(p) = D(p) + M(p) \tag{1}$$

The distance $D(p)$ is the orthogonal distance between the segment connecting the two end points and the point p of the polyline. $M(p)$ represents the measure at the point p. A node in the structure is created to represent a point p_i, which importance value is $M(p_i)$.

The polyline (p_1, p_n) will be processed as follows: If the node root is represented by p_k (a point on the polyline) having the highest importance value, the creation of other nodes follows an iterative process addressing all the points of the segments $[p_1, p_k]$ and $[p_k, p_n]$. To illustrate this process, we propose the following example on road risk analysis; we focus on the number of accidents recorded on road segments connecting ten cities represented by c_1 to c_{10} points (see Figure 3). Each segment carries a measure that represents the number of accidents reported on the segment connecting city c_i to city c_{i+1}.

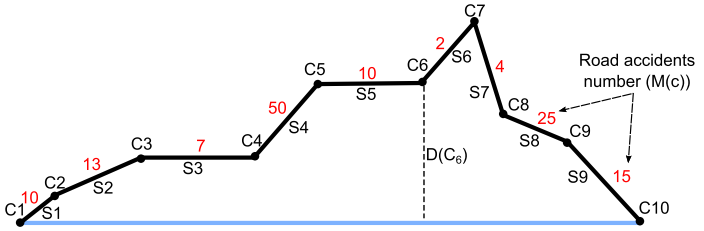


Fig. 3. Polyline (road) representation in SOLAP

We emphasize that in this case, the measures are associated with road segments, or the process requires their transposition to the endpoints constituting

these segments. To do this we propose that the measure of each point p_i is determined by the maximum value of the measure of the segments to which the point pi belongs: Let $M(p_i)$ be the measure at point p_i.

$$M(p_i) = Max(M([p_{i-1}, p_i]), M([p_i, p_{i+1}])$$

Furthermore, the values of the measures and distances such as identified have different domains. Indeed there is a significant difference between these two parameters. A normalization step is thus necessary, in order to make the values comparable to each other. To do this we will restrict values between 0 and 1. For each value V of a measure or a distance, its normalized value V' is calculated as follows:

$$V' = \frac{V - V_{min}}{V_{max} - V_{min}} \tag{2}$$

Thus, the BLG structure of the original polyline depicted in figure 3 is as shown in figure 4.

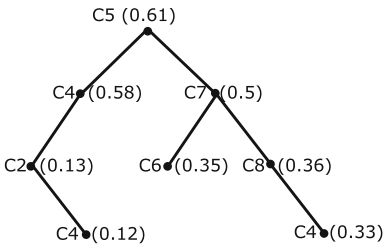


Fig. 4. The built SOLAP BLG tree structure

3.2 Proposed Generalization Process

The on-the-fly generalization process adapted to SOLAP context is guided by the SOLAP BLG tree structure according to the method described below.

Once these structures built, they are saved in a session work. When navigating between the different levels of detail, the on-the-fly generalization process is triggered to retrieve only the visible points in the required level of detail. The selection of these points is performed by comparing the importance values previously stored with a threshold value. The latter is determined by the visualization scale and other cartographic parameters that require the intervention of an expert cartographer. In the context of this work we used experimentally determined thresholds. Indeed, the threshold determines the tree traversal depth, by selecting only the nodes whose relevance value is greater than the threshold.

Measures Aggregation. The measures associated with the different objects are subject to an aggregate function that determines the measures of the resulting objects. This maintains the importance of the decision-making aspect of the different requested levels of detail. This aggregation function is developed according to the analyzed fact. For example for the analysis of the road risk phenomenon, the proposed aggregation function is the sum function, to preserve the information on the total number of accidents on the generalized object.

To aggregate polylines measures, we propose the creation of a data structure, containing the values of the measures associated with the different segments constituting the initial polyline depicted in Figure 3 as shown in the example of table 1.

Table 1. Data structure dedicated to measures storage

Segment	S1	S2	S3	S4	S5	S6	S7	S8	S9
Measure	10	13	7	50	10	2	4	25	15

During the generalization with the BLG tree, we obtain a new polyline where the segments S_i to S_k are removed and replaced by a new segment formed by the first point of S_i and the last point of S_k, to evaluate the measure associated with this new segment, one can read the above table and sum the measures corresponding to the segments from S_i to S_k.

To illustrate this process we will use the polyline shown in Figure 5, in each segment we took the values of measures in accordance with the table above. After simplification of the polyline we get two segments formed by the points C_1, C_5 and C_{10}, the points C_2 through $C4$ are deleted along with the points C_6 through C_9. The measure associated with the new segment $[C_1, C_5]$ represents the sum of the measures contained in the table (measure of the segments S_1 through S_4).

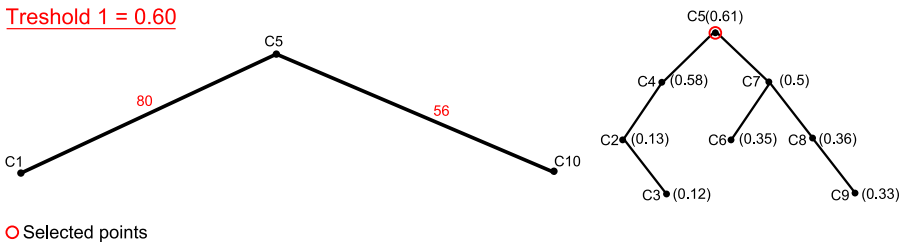


Fig. 5. SOLAP BLG tree generalization results

It is therefore clear that the displayed map will be simplified in order to highlight the information requested by the query.

4 Experimentation

The validation of our approach involves the construction of the proposed spatial structure SOLAP BLG tree used in in the implemented generalization process. We choose the road risk as a case study given its socio-economic impact worldwide. According to statistics from the World Health Organization [23], the road causes each year more than 1.2 million deaths and between 20 to 50 million wounded. In a previous work, we addressed this phenomenon by incorporating spatial information [24] [25].

Our tests are performed on vector spatial data. We used multiple softwares and hardawre ressources to implement our generalization prototype : (1) Oracle 11g Enterprise Edition as DBMS (Management System Database) via its component Oracle Spatial. (2) Oracle MapViewer for viewing the map of the analyzed area. (3) Oracle Weblogic Server on which MapViewers components are deployed. (4) Oracle Map Builder was used to load the geographical data in the DBMS and the construction of the map. (5) Oracle JDeveloper tool as a code editor.

Our experiments were performed on a data set that represents the road theme of Dar El Beida municipality in Algiers enriched by different measures representing the number of accidents recorded on the considered roads (Figure 6).

4.1 SOLAP BLG tree Test

To test the SOLAP BLG tree, we selected a road in Dar El Beida municipality. This road is shown in red in Figure 6. It includes 28 segments each one having a measure. Our generalization system simplifies this road at smaller scales, taking into account the decision aspect (measures).

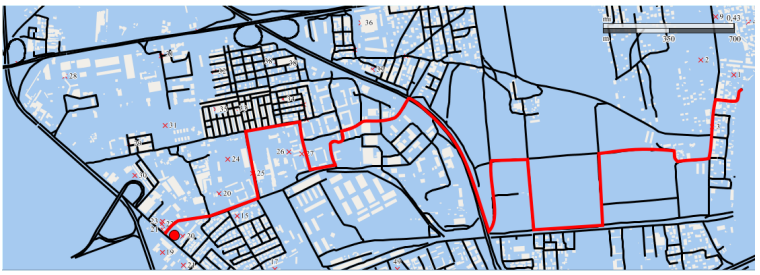


Fig. 6. Road network of Dar El Beida and the selected road

Figure 7 illustrates a detailed representation of the selected route; segments with the highest number of accidents are highlighted.

The SOLAP BLG tree corresponding to the selected route is shown in Figure 8. The root node contains the point p_{14}, which has the highest importance value. Points stored in the top levels are the points having the highest importance values.

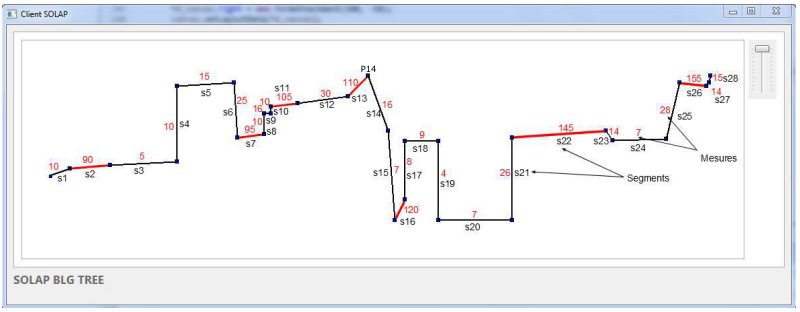


Fig. 7. Detailed description of the selected raod

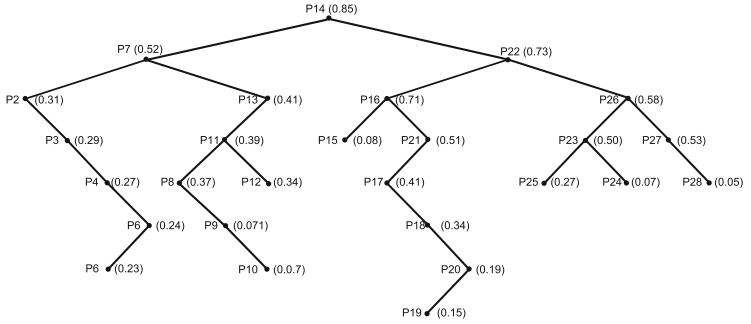


Fig. 8. SOLAP BLG tree corresponding to the selected road

Following the proposed approach, the generalization of the selected road allows to restore only the segments points visible at the required scale. From a more detailed scale, we can analyze the results obtained at different scales (see Figure 5). For example in scale 1: 5000 all relevant segments road except the

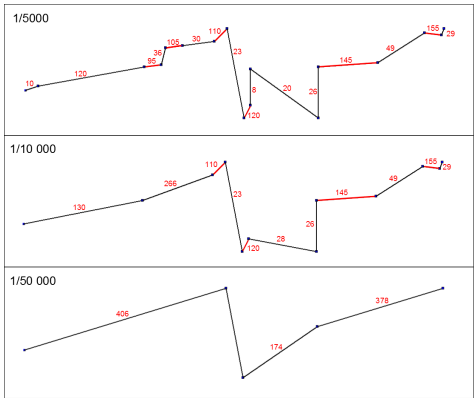


Fig. 9. Generalization results guided by SOLAP BLG tree

segment $S2$ are visible on the map, whereas, at the scale 1: 10,000, there are only four relevant segments and at the scale 1: 50,000 relevant segments are no longer visible.

Figure 10 shows a comparison between the results of the generalization with SOLAP BLG as part of this work and the results of the generalization with the classic BLG tree. We can see that at the same scale 1: 10 000, relevant road segments are visible in the case of SOLAP BLG tree (segments S13, S16, S22 and S26), while they are no longer in the case of classical BLG tree, despite their decision relevance, hence the importance of generalization with SOLAP BLG tree in the SOLAP context.

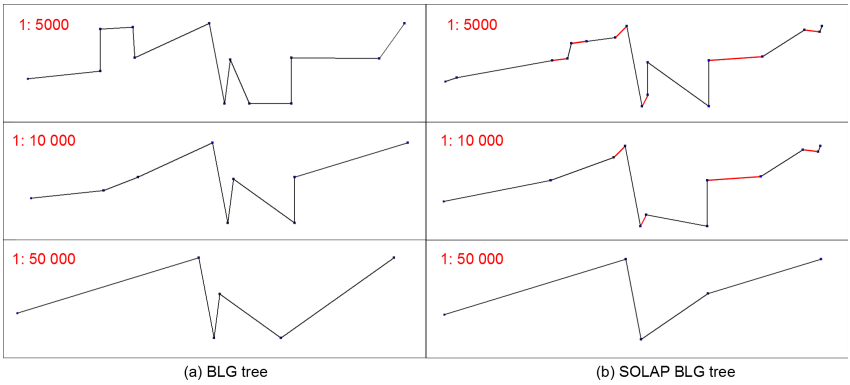


Fig. 10. Comparaison between the results of generalization by the SOLAP BLG tree and the BLG tree

5 Conclusions and Future Issues

This paper presented an on-the-fly generalization approach adapted to SOLAP applications. This process is intimately linked to the highly interactive applications in cartography such as web mapping, mobile mapping and SOLAP applications. But this latters, require the simultaneous consideration of cartographic and decisional aspects by integrating the measure in the process.

Our proposed approach is based on SOLAP BLG tree, it consists on adapting BLG tree structure, initially dedicated to cartographic generalization, to SOLAP. It focuses on linear objects (roads in our case study) and integrates the measure in order to adapt the level of detail that meets the decision-makers needs.

To validate our approach, we chose the road risk phenomenon as analysis subject, this is particularly due to its worldwide socio-economic impact. In addition, the use of map in the analysis of such phenomenon is of major interest for decision-makers because it is closely related to geographic information represented by the road object and the locality to which it belongs. In our experiments, we have highlighted the contribution of the proposed structure in the context

of SOLAP through the various implemented functions such as importance function, and whose application has allowed preserving measures while providing cartographic perceptibility.

As future issues, we suggest : (1) improving the current solution by adapting the other generalization operators (as is smoothing, displacement, typification, exaggeration, etc.) to SOLAP applications. This will allow generating a better quality of maps and hence, improve the decision making process. (2) Adapt the generalization process to SDW by using another generalization approach, for example, the one, based on rapid generalization algorithms and (3) elaborate a comparative study between the implemented approaches according to some defined criteria in order to assess their effectiveness in a given context of use.

References

1. Chaudhuri, S., Dayal, U.: An overview of data warehousing and olap technology. ACM Sigmod Record 26(1), 65–74 (1997)
2. Kimball, R., Ross, M., et al.: The data warehouse toolkit: the complete guide to dimensional modelling. Wiley, New York (2002)
3. Thomsen, E.: OLAP solutions: building multidimensional information systems. John Wiley & Sons (2002)
4. Rivest, S., Bédard, Y., Proulx, M.J., Nadeau, M., Hubert, F., Pastor, J.: Solap technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. ISPRS Journal of Photogrammetry and Remote Sensing 60(1), 17–33 (2005)
5. Malinowski, E., Zimányi, E.: Advanced data warehouse design: from conventional to spatial and temporal applications. Springer Science & Business Media (2008)
6. Bédard, Y.: Spatial olap. In: Forum Annuel sur la RD, Géomatique VI: Un Monde Accessible, pp. 13–14 (1997)
7. Bimonte, S., Bertolotto, M., Gensel, J., Boussaid, O.: Spatial olap and map generalization: Model and algebra. International Journal of Data Warehousing and Mining (IJDWM) 8(1), 24–51 (2012)
8. Bereuter, P., Weibel, R.: Algorithms for on-the-fly generalization of point data using quadtrees. In: Proceedings AutoCarto 2012 (2012)
9. Weibel, R., Burghardt, D.: Generalization, on-the-fly. In: Encyclopedia of GIS, pp. 339–344. Springer (2008)
10. van Oosterom, P., Meijers, M.: Towards a true vario-scale structure supporting smooth-zoom. In: Proceedings of the 14th ICA/ISPRS Workshop on Generalisation and Multiple Representation, Paris, vol. 48 (2011)
11. van Oosterom, P., Meijers, M., Stoter, J., Šuba, R.: Data structures for continuous generalisation: tgap and ssc. In: Abstracting Geographic Information in a Data Rich World, pp. 83–117. Springer (2014)
12. Sarjakoski, L.: Conceptual models of generalisation and multiple representation. In: Generalisation of Geographic Information: Cartographic Modelling and Applications. Elsevier, Amsterdam (2007)
13. Gaffuri, J.: Généralisation automatique pour la prise en compte de thèmes champ: le modèle GAEL. PhD thesis, Université Paris-Est (2008)
14. Stanislawski, L.V., Buttenfield, B.P., Bereuter, P., Savino, S., Brewer, C.A.: Generalisation operators. In: Abstracting Geographic Information in a Data Rich World, pp. 157–195. Springer (2014)

15. Lehto, L., Sarjakoski, L.T.: Real-time generalization of xml-encoded spatial data for the web and mobile devices. International Journal of Geographical Information Science 19(8-9), 957–973 (2005)
16. Foerster, T., Lehto, L., Sarjakoski, T., Sarjakoski, L.T., Stoter, J.: Map generalization and schema transformation of geospatial data combined in a web service context. Computers, Environment and Urban Systems 34(1), 79–88 (2010)
17. GiMoDig: Geospatial info-mobility service by real-time data-integration and generalisation (2001)
18. Lejdel, B., Kazar, O.: Genetic agent approach for improving on-the-fly web map generalization. CoRR (2012)
19. Van Oosterom, P.: The reactive-tree: A storage structure for a seamless, scaleless geographic database. In: Autocarto-Conference. ASPRS American Society for Photogrammetry and Remote Sensing, vol. 6, pp. 393–393 (1991)
20. Van Oosterom, P.: Variable-scale topological data structures suitable for progressive data transfer: The gap-face tree and gap-edge forest. Cartography and Geographic Information Science 32(4), 331–346 (2005)
21. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica: The International Journal for Geographic Information and Geovisualization 10(2), 112–122 (1973)
22. Bereuter, P., Weibel, R.: Real-time generalization of point data in mobile and web mapping using quadtrees. Cartography and Geographic Information Science 40(4), 271–281 (2013)
23. Bencherif, H., Boubakour, F., Belkacem, N.: Les accidents de la route dans les médias de masse en algérie. du traitement de l'information à sa diffusion. Communication Information Médias Théories Pratiques 30(1) (2012)
24. Amieur, K.D., Frihi, I., Boukhalfa, K., Alimazighi, Z.: De la conception d'un entrepôt de données spatiales à un outil géo-décisionnel pour une meilleure analyse du risque routier (2013)
25. Derbal, K., Ibtissem, F., Boukhalfa, K., Alimazighi, Z.: Spatial data warehouse and geospatial decision making tool for efficient road risk analysis. In: 2014 1st International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), pp. 1–7. IEEE (2014)

QoS-Aware Web Services Selection Based on Fuzzy Dominance

Amal Halfaoui(✉), Fethallah Hadjila, and Fedoua Didi

Computer Science Departement,
Tlemcen University, Algeria
{a_halfaoui,f_hadjila,f_didi}@mail.univ-tlemcen.dz
http://www.univ-tlemcen.dz/

Abstract. The selection of an appropriate web service for a particular task has become a difficult challenge due to the increasing number of web services offering similar functionalities. Quality of web services (QoS) becomes crucial for selecting web services among functionally similar components. However, it remains difficult to select an interesting Web services from a large number of candidates with a good compromise between multiples QoS aspect. In this paper, we propose a novel concept based on dominance degree to rank functionally similar services. We rank Web services by using a fuzzification of Pareto dominance called Average-Fuzzy-Dominated-Score($AFDetS()$). We demonstrate the effectiveness of the $AFDetS$ through a set of simulations by using a real Dataset.

Keywords: Web service selection · dominance · Skyline · Ranking · QoS

1 Introduction

Nowadays, an increasing number of Web services is published and accessible over the web, they are designed to perform a specific task, which essentially consists of either altering the word state (e.g.,an on line shopping service) or returning some information to the user (e.g.,news Web service).

As the Web is populated with a considerable number of Web services, there exists a large number of service providers competing to offer the same functionality, but with different Quality Of Service(QoS) such as response time, price, etc. Consequently, QoS is thus a crucial criterion to select among functionally similar Web services.

Example. Consider a Web service for sending SMS, there are many Web services providing this functionality (e.g., Click Send, Inteltech, Etc.), but with different QoS. Table1 provides such functionality along with real QoS parameters taken from the publicly available Quality of Web services data.[1] Web services were obtained by using the keyword SMS which represents the tag associated to the functionality of the desired Web services. Each Web service has four *Qo*S parameters q_1, q_2, q_3 and q_4, says respectively Response Time, Throughput (i.e.,Total

[1] http://www.uoguelph.ca/~qmahmoud/qws

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 291–300, 2015.
DOI: 10.1007/978-3-319-19578-0_24

Table 1. A set of Sending SMS Web Services

Service	provider	operation	$q1(ms)$	$q2$(hits/sec)	$q3(\%)$	$q4(\%)$
S_1	acrosscommunications.com	SMS	113.8	5.2	81	84
S_2	sjmillerconsultants.com	SMS	179.2	0.7	65	69
S_3	webservicex.net	SendSMS	1308	6.3	67	84
S_4	webservicex.net	SendSMSWorld	3103	5.3	79.3	91
S_5	smsinter.sina.com.cn	SMSWS	751	6.8	64.3	87
S_6	sms.mio.it	SendMessages	291.07	5.2	53.6	84
S_7	www.barnaland.is	SMS	436.5	4.5	43.2	84
S_8	emsoap.net	emSoapService	424.54	4.3	11.9	80

Number of invocations/period of time) Reliability(Ratio: number of error messages/total messages) and Best Practices (the respect of the specifications). To select an adequate Web service, users need to examine all of them manually. The user may also face difficulties in balancing between different quality metrics. The skyline presents a good solution for reducing the number of candidate Web services [1],[2] and simplifying the process of selection as it overcomes the major limitation of the current approaches that require users to assign weights over different QoS attributes. The skyline is a subset of Web services that are not(Pareto) dominated by any other Web service. A Web service S_i is said to Pareto domine another Web service S_k if and only if S_i is better than or equal to S_k in all QoS parameters and better than S_k in at least on one QoS parameter.

According to our example (Table1), the service S_1 dominates S_6,S_7,S_8. Services S_1,S_3,S_4,S_5 belong to the skyline and they are no comparable between them. We can remark that computing skyline reduce the candidates services, in our example we eliminates 50% of the candidate services. However, it remains a challenge to compute skylines in high dimensional data [3],[4]. In addition to that, on the report of [5] the authors show that the skyline may lose some interesting Web services like S_6 which is dominated by S_4 while S_4 is the worst service in term of response time, however S_6 has a good response time and is closer to S_4 on the other QoS parameters.

Motivated by this, we propose an extension of Pareto dominance relationship called Averaged-Fuzzy-Dominated-score $AFDetS()$ to associate a score to each service and rank them, We also propose a comparison between the dominated-score $AFDetS$ and Dominating score used in [5] and confirm that the use of the Dominated score is more interesting than the Dominating score in Ranking service, this fact is also confirmed in [19]. The rest of the paper is organized as follows. In the next section, we discuss related work. In Section 3, we provide the formal definition of $AFDetS$ and show it application on our example Table1. Section 4 presents the results of our experimentation. Finally, section 5 gives conclusions and an outlook on possible continuations of our work.

2 Related Work

A lot of efforts have been devoted to the problem of QoS-aware Web service selection. Some of them use the linear programming technique [7], [8]. Linear programming techniques are used in [7] to find the optimal selection of component services and gives an extensible model to evaluate the QoS parameters, Linear programming techniques are extended in [8] to include local constraints. Others work use combinatorial model and graph model [9] where the authors use heuristic algorithm to solve the problem of service selection with multiple QoS constraints. In [10] the authors present a selection algorithm to evaluates multiples QoS based on an ontology. Nevertheless, the majority of these approaches are more suitable for limited number of Services(the selection process has an exponential space complexity) and limited number of QoS, especially when the users has to assign weights on QoS attributes.

In recent research, the skyline paradigm is introduced as a good and efficient mechanism to reduce the number of service candidates and simplify the process of selection. The idea of skyline comes from the old research like contour problem, maximum vector and convex hull and was introduced into databases by Borzsonyi [11] who develops three algorithms: BNL, DC and B-tree, this leads to develop and ameliorate several other algorithms like SFS [12], SaLSa [13], Zorder, [14] and NN[3]. Some of these algorithms exploit index structures like [14],[3] to enhance the skyline computation process. However, the size of skyline increases under a high number of QoS and sometimes privileges Web services with bad compromise between QoS.

To handle the problem of large skyline, some works combine the advantage of the skyline and ranking and define variants of skyline like [1],[15],[16] and [17]. In [15] the authors present skyline frequency concept which is the number of subspaces where a point p is skyine, however this lead to calculate skyline of all subspaces and results in a hight computational time, further more authors introduce an approximate algorithm to reduce the computation space. In [1] Chan et al. present the notion of k-dominance which relax the pareto dominance to a subset of k parameters, however There exists cyclic dominance relationship (CDR) which leads to the loss of skylines in addition k-dominance often returns an empty set. In [17] lin et al. propose top-*krepresentative* skyline but this method is more suitable for anti-correlated data [18] in addition to that, k-representative skyline is considered as NP-hard for more than three dimensional dataset. In [16] the authors present the skyline graph which maps the dominance of different skyline subspaces into a weighted directed graph and use link-based techniques to rank skyline, however, the problem of dominance on a large space is still solved. These approaches rely on Pareto dominance relationship thus, they don't consider or privilege services with a good compromise between parameters, this drawback can be solved by the fuzzification of Pareto dominance in order to rank incomparable services.

The Fuzzy dominance was used in databases community like [20] the authors show the goal of fuzzification of the concept of Pareto dominance and it application in Evolutionary Multiobjectif Optimization. Other works use this principle

and applied it in Genetic or particle Swarm Algorithm. In service computing community, [5] use the fuzzy-dominance and propose the α-dominance to rank Web service based on QoS parameters and associates the fuzzy-dominating score to Web services.

Like mentioned in [20] the measures between two vectors a, b "a dominates b by degree α" and "a is dominated by b to degree α" is not symmetric, In addition to that, in [19] the authors demonstrates that the use of the dominated measure is more efficient in selecting the top-k services than the dominating measure. Our work is close to [5]. However, [5] use fuzzy-dominating relationship to compare the services instead of use fuzzy-dominated measure in ranking services. According to these observations, we define the Fuzzy Dominated relationship $Fdet$ and the Average Fuzzy dominated Score $AFDetS()$. The next section presents the definition of this concept and it utilization in our context.

3 Problem Formalization

In this section, we are going to study the fuzzification of the Pareto dominance relation , and show it application on our example (Table1). To allow for a uniform measurement of Web Services, we first normalize the different QoS value in the range [0,1].

3.1 Normalization of QoS Parameters

let be S a set of similar functionally services $S = S_1, .., S_n$. Suppose that we have R quantitative Q_oS values for a service S_i. we use the vector $Q(S_i) = \{Nq_1(S_i), .., Nq_r(Si)\}$ to represent the Q_oS attributes of a service S_i where the function $Nq_k(S_{ij})$ represent the k-th Normalized quality attribute of S_i. We convert the negative attributes(time, cost) into positive attributes by multiplying their values by -1 so that the higher value is the higher quality. We normalize the different QoS values in the range [0, 1], as follow :

$$Nq_k(S_i) = \frac{q_k(S_i) - Qmin(q_k)}{Qmax(q_k) - Qmin(q_k)} \tag{1}$$

Where $Nq_k(S_{ij})$ is the normalized QoS value of the Web service S_{ij} on the QoS parameter q_k and $Qmin(q_k)$ (resp.$Qmax(q_k)$ is the minimum (resp. maximum) value of the QoS parameter q_k. Table2 shows the QoS values of Web services example of Table 1 after normalization.

3.2 Fuzzification of Pareto Dominance Relation

Services of the same functionality differ only in term of Qos. Like mentioned above, the skyline consists of the set of points which are not pareto dominated by any other.

Table 2. Web Services with Normalized *QoS*

Web service	$Nq1$	$Nq2$	$Nq3$	$Nq4$
s1	1	0.74	1	0.68
s2	0.98	0	0.77	0
s3	0.60	0.92	0.80	0.68
s4	0	0.75	0.98	1
s5	0.79	1	0.76	0.82
s6	0.94	0.74	0.60	0.68
s7	0.89	0.62	0.45	0.68
s8	0.90	0.59	0	0.50

Definition 1 (*Pareto Dominance.*) *Let S_i and S_j be two Web services,Given a set of d QOS parameters $Q = \{q_1, ..., q_d\}$, We say that S_i dominates S_j denoted by $S_i \succ S_j$, iff $\forall q_k \in Q$, $q_k(S_i) \geq q_k(S_j)$ and $\exists q_t \in Q$, $q_t(S_i) > q_t(S_j)$.*

Pareto dominance does not differentiate between Web services with good compromise and those with bad compromise, to clarify this, let us return to our example (Table2) and consider S_4 and S_5, in fact neither S_4 dominates S_5 nor S_5 dominates S_4, the two services are incomparable and belong to the skyline because S_4 is better than S_5 in q_3 and q_4, and S_5 is better than S_4 in q_1 and q_2. However we can consider that S_5 is better than S_4 since $q_1(S_5) = 0.79$ is much higher than $q_1(S_0) = 0$. In addition to that, $q_3(S_5) = 0.76$ and $q_4(S_5) = 0.82$ are almost close to (respectively) $q_3(S_4) = 0.98$ and $q_4(S_4) = 1$. For this reason, it is interesting to fuzzify the Pareto dominance. The goal of the fuzzification of Pareto dominance is to allow a practically usable numerical comparison between two service and express the extent to which a Web service (more or less) is dominated by another one.

To compute the Fuzzy dominance degrees it's important to distinguish between the measure of two concepts : the dominating score and the dominated Score between two service S_i and S_j. The first one express the degree to which S_i dominates S_j and the second express the degree to which S_i is dominated by Sj and the measure of dominance is not symmetric. We will use in our work the concept of dominated relation. We define bellow the fuzzification of the dominated relation.

Definition 2 (*Fuzzy-Dominated Score.*) *let be S a set of functionally similar services, S_i and $S_j \in S$. Let $Q = \{q_1, ..., q_d\}$ be a vector of d QoS parameters. First we define the monotone comparison function $\mu_{\varepsilon,\lambda}$ to express the degree to which u is dominated by v, where u represent $q_k(s_i)$ and v represent $q_k(s_j)$ as follow:*

$$\mu_{\varepsilon,\lambda}(u,v) = \begin{cases} 0 & \text{if } (u-v) \geq \varepsilon \\ |u-v-\varepsilon| \,/\, |\lambda+\varepsilon| & \text{if } \lambda+\varepsilon \leq (u-v) < \varepsilon \\ 1 & \text{if } (u-v) < \lambda+\varepsilon \end{cases} \tag{2}$$

Where $\varepsilon, \lambda \in [-1, 0], \varepsilon + \lambda \geq -1$
Then, we define the Fuzzy-Dominated score $FDet(S_i, S_j)$ *to express the degree to witch* S_i *is dominated by* S_j *as follow:*

$$FDet(s_i, s_j) = \frac{1}{d} \sum_{k=1}^{d} \mu_{\lambda,\varepsilon}(q_k(s_i), q_k(s_j))) \tag{3}$$

Let us reconsider our example and compare Web services S_4 and S_5 by using $FDet()$, with $\varepsilon = -0.1$ and $\lambda = -0.2$ we have $FDet(S_4, S_5) = 0.5$ and $FDet(S_5, S_4) = 0$ this mean that S_5 is not fuzzy dominated by S_4 and is little more better than S_4. This concept gives a good compromise between QoS. In fact, this is more expressing than S_4 and S_5 not comparable by Pareto dominance. In what follows, we use the $FDet()$ to rank Web services

Definition 3 (*Averaged-Fuzzy-Dominated-Score.*) *In order to rank a Web service* S_i *in it class* S*, we first, make pairwise comparison with the other services and associate it a score by:*

$$AFDetS(S_i) = \frac{1}{|S| - 1} \sum_{j=1, i \neq j}^{n} FDet(S_i, S_j) \tag{4}$$

Then, we retain service with lower $AFDetS()$ *on a higher ranking position*

The Table3 show the services of our example (Table 1) after computing $AFDetS$ score and ranking with $\varepsilon = 0$ and $\lambda = -0.2$

Table 3. Services'Rank according to $AFDetS()$

Rank Web service	AFDedS()	$Nq1$	$Nq2$	$Nq3$	$Nq4$
s1	0,071	1	0,74	1	0,68
s5	0,107	0,79	1	0,76	0.82
s6	0,143	0,94	0,74	0,60	0,68
s3	0,25	0,60	0,92	0,80	0,68
s7	0,286	0,89	0,62	0,45	0,68
s4	0,312	0	0,75	0,98	1
s8	0,393	0,90	0,59	0	0,50
s2	0,571	0,98	0	0,77	0

We can observe that the top service is S_1 which is better than the others in q_1, q_2 and has a good value in the other QoS parameters. We remark that services that have some $QoS = 0$ are at the bottom of the ranking. Let us consider S_6 and S_4, according to the result provided by Pareto dominance S_4 belong to the skyline, but S_6 does not, however S_4 have the worst response time(q_1)) and S_6 has a good compromise between QoS parameters. According to (Table3: Fuzzy-Dominated Score)S_4 was downgraded to the Rank 7, On the other hand, the

Service S_6 which has a good compromise between QoS parameters was set up to the 3rd rank.

From this result, we confirm that the use of $Fed()$ can give more interesting results in term of balanced of QoS than the other approaches.

4 Experimental Evaluation

In order to evaluate and prove the effectiveness of our approach, we compare the result of using Fuzzy-Dominated with the Fuzzy-Dominating score. For this purpose, we implement the function fuzzy-dominating proposed in [6] and termed it $AFDingS$ and compare it to our Approach $AFDetS$. All the experiments are conducted on the same software and hardware, which were Intel i3-2365M CPU @ 1.40GHz 4 processors, 4.0GB of RAM, Ubuntu 13.10, Netbeans 7.4. Several simulations have been made by varying the parameters:

- ε, λ,
- d:number of QoS parameter,
- n:number of services of the same class S.

For each simulation we take the Top-5 services generated by the algorithms $AFDetS$ and $AFDingS$ and compare them. Different Services' subsets were taken from the real QoS dataset provided by [23]. The dataset includes informations about 2507 real-world web services. Each service comprise measurement of nine QoS parameters. The service name and its WSDL address are also included in the dataset. We group functionally similar Services into clusters, for example the cluster "sms" (sending sms) contains 30 real services. The cluster "search"(ie. Search Engine Web services such as Google Search,Amazone, etc.) contain 92 services.

a-Varying ε and λ: We present below two scenarios (Table4) and (Table5) by varying ε and λ on a set of 30 services belonging to the class SMS. Each service has 4 QoS parameters.

Table 4. Top-5 Services Rank according to $AFDingS$,$AFDetS()$ with $\varepsilon = 0, \lambda = -0.2$

Top-5 AFDingS			Top-5 AFDetS		
Si	$AFDingS$	$Qos(q_1, q_2, q_3 q_4)$	Si	$AFDetS$	$Qos(q_1, q_2, q_3 q_4)$
S5	0.566	[0.787, 1.0, 0.758, 0.818]	S12	0.071	[1.0, 0.738, 1.0, 0.682]
S4	0.551	[0.0, 0.754, 0.975, 1.0]	S5	0.107	[0.787, 1.0, 0.758, 0.818]
S12	0.529	[1.0, 0.738, 1.0, 0.682]	S6	0.143	[0.941, 0.738, 0.603, 0.682]
S30	0.423	[0.6, 0.918, 0.797, 0.682]	S30	0.25	[0.6, 0.918, 0.797, 0.682]
S6	0.329	[0.941, 0.738, 0.603, 0.682]	S7	0.286	[0.0, 0.754, 0.975, 1.0]

Table 5. Top-5 Services according to $AFDingS$, $AFDetS()$ with $\varepsilon = -0.1, \lambda = -0.2$

Top-5 AFDingS			Top-5 AFDetS		
Si	$AFDingS$	$Qos(q_1, q_2, q_3, q_4)$	Si	$AFDetS$	$Qos(q_1, q_2, q_3 q_4)$
S4	0.443	[0.0, 0.754, 0.975, 1.0]	S5	0.0	[0.787, 1.0, 0.758, 0.818]
S5	0.421	[0.787, 1.0, 0.758, 0.818]	S12	0.036	[1.0, 0.738, 1.0, 0.682]
S12	0.036	[1.0, 0.738, 1.0, 0.682]	S6	0.107	[0.941, 0.738, 0.603, 0.682]
S30	0.321	[0.6, 0.918, 0.797, 0.682]	S30	0.143	[0.6, 0.918, 0.797, 0.682]
S6	0.223	[0.941, 0.738, 0.603, 0.682]	S7	0.25	[0.892, 0.623, 0.453, 0.682]

We can observe from the results on (Table 4) and (Table 5) that the ranking given by $AFDetS$ is more interesting than the one given by $AFDingS$ even if we vary ε and λ the top-1 is always better according to $AFDetS$. The service S4 (Table 5)is the top-1 according to $AFDing$ while it does not belong to the top-5 according to $AFDetS$ because of its bad first criterion value. We can say that $AFDetS$ favors services with good value in all parameters and discards services with worst values in some QoS parameters even if the others are good.

b-Varying *d* and *n*: We present below two scenarios by varying d from 7 to 9 on a set of 92 services belonging to the class search. We fixed $\varepsilon = -0.1$ and $\lambda = -0.2$. The result of the top-5 services provided by $AFDtingS$ and $AFDetS$ approach are shown in (Table6) and (Table7).

Table 6. Top-5 Services($AFDingS()$ Vs. $AFDetS()$) with $d = 7$

	Si	$Score$	$Qos(q_1, q_2, q_3 q_4, q_5, q_6, q_7)$
AFDingS	S70	0.409	[0.183, 0.904, 0.618, 0.964, 0.767, 1 , 0.815]
	S30	0.388	[0.164, 0.904, 1 , 0.964, 0.767, 1 , 0.815]
	S24	0.385	[0.005, 1 , 0.829, 1 , 0.767, 1 , 0.667]
	S72	0.381	[0.474, 0.795, 0.260, 0.807, 0.767, 0.667, 0.815]
	S16	0.365	[0.003, 1 , 0.419, 1 , 1 , 0.667, 0.111]
AFDetS	S30	0.005	[0.164, 0.904, 1 , 0.964, 0.767, 1 , 0.815]
	S52	0.006	[0.016, 0.819, 0.955, 0.94, 0.767, 1 , 0.667]
	S24	0.006	[0.005, 1 , 0.829, 1 , 0.767, 1 , 0.667]
	S70	0.008	[0.183, 0.904, 0.618, 0.964, 0.767, 1 , 0.815]
	S45	0.022	[0.042, 0.831, 0.382, 0.940, 0.767, 1 , 0.667]

From (Table6), we can observe that the ranking given by $AFDetS$ is more interesting than the one given by $AFDingS$. The top-1($AFDetS$) is the service S30. This latter has better value than the top-1($AFDingS$) on q_3. Moreover, service S30 is close to service S7 on q_1 parameter. We can remark that the service S16 is included into top-5($AFDingS$) while it does not belong to the top-5($AFDetS$) because of its bad values on q_3 and q_7. In fact,it is replaced by service S45 witch has a good compromise between its QoS parameters.

Table 7. Top-5 Services($AFDingS()$ Vs. $AFDetS()$) with $d = 9$

	Si	$Score$	$Qos(q_1, q_2, q_3q_4, q_5, q_6, q_7, q_8, q_9)$
AFDingS	S24	0.397	[0.050, 1 , 0.829, 1 , 0.767, 1 , 0.667, 0.004, 0.958]
	S16	0.366	[0.003, 1 , 0.419, 1 , 1 , 0.667, 0.111, 0.030, 0.358]
	S60	0.344	[0.016, 0.988, 0.955, 1 , 0.333, 1 , 0.259, 0.008, 0.337]
	S55	0.328	[0.179, 0.916, 0.244, 0.976, 0.767, 1 , 0.815, 0.066, 0.800]
	S70	0.318	[0.183, 0.904, 0.618, 0.964, 0.767, 1 , 0.815, 0 , 0.021]
AFDetS	S24	0.006	[0.050, 1 , 0.829, 1 , 0.767, 1 , 0.667, 0.004, 0.958]
	S45	0.018	[0.042, 0.831, 0.382, 0.940, 0.767, 1 , 0.667, 0.030, 0.937]
	S55	0.018	[0.179, 0.916, 0.244, 0.976, 0.767, 1 , 0.815, 0.066, 0.800]
	S52	0.024	[0.016, 0.819, 0.955, 0.940, 0.767, 1 , 0.667, 0.017, 0.105]
	S30	0.027	[0.064, 0.904, 1 , 0.964, 0.767, 1 , 0.815, 0.092, 0.053]

Let us consider now the ranking with $d = 9$ (Table7). The two ranking methods have the same top-1 (service S24). However, the other services given by $AFDetS$ are different from those provided by $AFDingS$. The service S16 and the service S70 witch belong to (top-5($AFDingS$)) are discarded by $AFDetS$ from the top-5 because they contains some bad values (close /or equal to 0) on some Qos criteria. This two services are replaced by respectively the service S45 and the service S30 by the $AFDetS$ approach, we can remark that these two services present a good compromise between their QoS parameters.

5 Conclusion

In this paper, we have presented an approach for ranking QoS-based-Web services. We have presented a fuzzification of the Pareto-dominance and introduced the concept $AFDetS$ which associates a score to a service according to the Fuzzy dominated relation. We demonstrate that the fuzzy dominated concept can offer an alternative to compare services when they are non comparable with pareto dominance. Experimental results show that the proposed approach is effective in comparison with the Fuzzy Dominating ranking. For future work, we can use this concept for the web service composition.

References

1. Chan, C.-Y., Jagadish, H.V., Tan, K.-L., Tung, A.K.H., Zhang, Z.: On high dimensional skylines. In: Ioannidis, Y., et al. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 478–495. Springer, Heidelberg (2006)
2. Lee, J., You, G., Hwang, S.: Personalized top-k skyline queries in high-dimensional space. Information Systems 34(1), 45–61 (2009)
3. Kossmann, D., Ramsak, F., Rost, S.: Shooting stars in the sky: an online algorithm for skyline queries. In: Proceedings of the 28th International Conference on Very Large Data Bases 2002, pp. 275–286. VLDB Endowment, Hong Kong (2002)

4. Papadias, D., et al.: Progressive skyline computation in database systems. ACM Transactions on Database Systems (TODS) 30(1), 41–82 (2005)
5. Benouaret, K., Benslimane, D., HadjAli, A.: On the use of fuzzy dominance for computing service skyline based on qos. In: ICWS 2011, pp. 540–547 (2011)
6. Benouaret, K., Benslimane, D., Hadjali, A.: A fuzzy framework for selecting top-k Web services compositions. Applied Computing Review (2011)
7. Zeng, L., Benatallah, B., Ngu, A.H.H., Dumas, M., Kalagnanam, J., Chang, H.: Qos-aware middleware for web services composition. IEEE Trans. Software Eng. 30(5), 311–327 (2004)
8. Ardagna, D., Pernici, B.: Adaptive service composition in flexible processes. IEEE Trans. Software Eng. 33(6), 369–384 (2007)
9. Yu, T., Zhang, Y., Lin, K.-J.: Efficient algorithms for web services selection with end-to-end qos constraints. TWEB 1(1) (2007)
10. Wang, X., Vitvar, T., Kerrigan, M., Toma, I.: A qos-aware selection model for semantic web services. In: Dan, A., Lamersdorf, W. (eds.) ICSOC 2006. LNCS, vol. 4294, pp. 390–401. Springer, Heidelberg (2006)
11. Borzsonyi, S., Kossmann, D., Stocker, K.: The Skyline Operator. In: Proceedings of the 17th International Conference on Data Engineering 2001, pp. 421–430. IEEE Computer Society (2001)
12. Chomicki, J., Godfrey, P., Gryz, J., Liang, D.: Skyline with Presorting. In: Proc. 19th IEEE Intl Conf. Data Eng. (ICDE), pp. 717–816 (2003)
13. Bartolini, I., Ciaccia, P., Patella, M.: Efficient Sort-Based Skyline Evaluation. ACM Trans. Database Systems 33(4), 1–45 (2008)
14. Lee, K.C.K., Zheng, B., Li, H., Lee, W.-C.: Approaching the skyline in z order. In: VLDB, pp. 279–290 (2007)
15. Chan, C.Y., et al.: Finding k-dominant skylines in high dimensional space, pp. 503–514 (2006)
16. Vlachou, A., Vazirgiannis, M.: Ranking the sky: Discovering the importance of skyline points through subspace dominance relationships. Data and Knowledge Engineering 69(9), 943–964 (2010)
17. Lin, X., et al.: Selecting stars: The k most representative skyline operator, pp. 86–95 (2007)
18. Alrifai, M., Skoutas, D., Risse, T.: Selecting skyline services for qos-based web service composition. In: WWW, pp. 11–20 (2010)
19. Skoutas, D., Sacharidis, D., Simitsis, A., Sellis, T.: Ranking and clustering web services using multi-criteria dominance relationships. IEEE Trans. on Services Computing (2010)
20. Koppen, M., Vicente Garcia, R.: A fuzzy scheme for the ranking of multivariate data and its application. In: Proceedings of the 2004 Annual Meeting of the NAFIPS (CD-ROM), Banff, Alberta, Canada, pp. 140–145 (2004)
21. Köppen, M., Vicente-Garcia, R., Nickolay, B.: Fuzzy-pareto-dominance and its application in evolutionary multi-objective optimization. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 399–412. Springer, Heidelberg (2005)
22. Köppen, M., Veenhuis, C.: Multi-objective particle swarm optimization by fuzzy-Pareto-dominance meta-heuristic. Int. J. Hybrid Intell. Syst. 3(4), 179–186 (2006)
23. Al-Masri, E., Mahmoud, Q.H.: Investigating web services on the world wide web. In: WWW, pp. 795–804 (2008)

Information Technology: Recommender Systems and Web Services

A Hybrid Model to Improve Filtering Systems

Kharroubi Sahraoui[1(✉)], Dahmani Youcef[2], and Nouali Omar[3]

[1] National High School of Computer Science E.S.I, and Ibn Khaldoun University Tiaret, Tiaret, Algeria s_kharoubi@esi.dz
[2] Department of Computer Science, Ibn Khaldoun University, Tiaret, Algeria
[3]
dahmani_y@yahoo.fr
[4] Basic Software Laboratory, C.E.R.I.S.T, Ben Aknoun, Algeria
o_nouali@cerist.dz

Abstract. There is a continuous information overload on the Web. The problem treated is how to have relevant information (documents, products, services etc.) at time and without difficulty. Filtering system also called recommender systems have widely used to recommend relevant resources to users by similarity process such as Amazon, MovieLens, Cdnow etc. The trend is to improve the information filtering approaches to better answer the users expectations. In this work, we model a collaborative filtering system by using Friend Of A Friend (FOAF) formalism to represent the users and the Dublin Core (DC) vocabulary to represent the resources "items". In addition, to ensure the interoperability and openness of this model, we adopt the Resource Description Framework (RDF) syntax to describe the various modules of the system. A hybrid function is introduced for the calculation of prediction. Empirical tests on various real data sets (Book-Crossing, FoafPub) showed satisfactory performances in terms of relevance and precision.

Keywords: Recommender systems · Resource description framework · Dublin core · FOAF · Semantic

1 Introduction

The multiplicity of the services offered via the Web excites the Net surfers to expose and communicate an enormous traffic of data of various formats. The gigantic mass of existing information and the speed of its instantaneous production triggers the problem of informational overload. This phenomenon known under the name big data imposes multiple difficulties such as management, storage, the control and the security of circulated data. On the other hand, the access to relevant information in time is a major occupation of the developers and users, in spite of his availability it is lost in the mass. The performances of the existing tools degrade when we handle large volume of data, more precisely the search engines are involved by this phenomenon in terms of recall and precision as well as the process of the indexing. Our work is more particularly listed under filtering information tab, specifically custom filtering in order to submit

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 303–314, 2015.
DOI: 10.1007/978-3-319-19578-0_25

the useful information to the users. Many commercial and educational sites are based on the filtering algorithms to recommend their products such as the Amazon, Movielens, Netflix, EducationWorld etc [5]. Filtering systems (FS), known as "recommender systems", have become essential with the increasing variety of web resources such as news, games, videos, documents or others [10]. The majority of the recent FS explores semantic information and share the metadata of the resources in order to improve the relevance factor[8]. Additionally, another type of these systems is based on ontology for conceptualizing and valorising the application domain, which makes it possible to increase their performances [1]. However, FS suffer from some common weaknesses, such as cold start, sparsity and scalability. In our study, we adopted the RDF model to represent all elements of the system with an open and interoperable manner. With the formalism Friend Of A Friend (FOAF), we weighted the attributes of the user profiles in order to gather them by degree of similarity. In addition, the items of system are represented by the Dublin Core vocabulary (DC) in RDF model to describe the web resources formally. These two formalisms that are recommended by W3C ensure interoperability and easy integration of the data. This approach allowed us to avoid focusing the approaches on a specific and closed field, and treats all kinds of resource using the URI and namespace clauses. The rest of the paper is organized as follows, we will briefly review the various forms of FS in section 2. The section 3 presents the details of our proposal. The results of experiments followed by discussions were exposed in section 4. In the end, we conclude our work with a conclusion and perspective.

2 State of the Art

The number of Internet users has now reached 38.8% of the world population in 2013 against 0.4% in 1995 according to statistics provided by ITU *(http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx)*. On the other hand, resources called commonly items occur at an incredible speed either by users or companies. Current tools are not consistent with this huge volume of data in order to analyze, control or have relevant information at time. The birth of FS is used to manage information overload by filtering [3,8]. Items can be extremely varied DVDs, books, images, web pages, restaurants ... etc. These systems are now increasingly present on the web and certainly will become essential in the future with the continuous increase of data [12]. According to how to estimate the relevance, researchers classify recommendation algorithms into three main approaches: content-based, collaborative and hybrid [4]. In the first approach, the system will support the content of the thematic items "documents" to compare them with a user profile, itself consists of topics explaining his interests, that is to say, the system compares the document themes with those of the profile and decides if the document is recommended or rejected according to the threshold of satisfaction function [17]. In the second approach, also known as social, the system uses the ratings of certain items or users and in order to recommend them to other users through the application of similarity process and without it being necessary

to analyze the content of items [2], in this approach, there are two main techniques which builds on memory-based algorithms, that operates a portion or all of the ratings to generate a new prediction [12] and which is founded on the model-based algorithms to create a descriptive model of the user so, estimate the prediction. The collaborative approaches are widely adopted in recommender systems such as Tapestry [4] GroupeLens [15], Amazon, Netflix ... etc. The hybrid methods operate to attenuate the insufficiencies of each of the two previous approaches by combining them in various manners. Recently, a new generation of FS boosted by semantic web formalisms or adaptable to contexts that uses a taxonomies or ontologies [13]. Commonly, these systems have shortcomings that prevent the recommendation process and degrade their performances, like the effect of the funnel where the user does not profited from the innovation and diversity of the items recommended in content-based filtering; the scalability where the system handles a large number of users and items online what makes difficult to predict in time; the sparsity problem, where there's a lack of sufficient evaluations to estimate the prediction well as the problem of the cold start to a user and/or item lately integrated into the system [11]. In this paper, we will extend the filtering systems in an open and interoperable specification, each component of the system is formalized by an appropriate RDF vocabulary. The following section explains the basic concepts of this specification.

3 Proposed Approach

Our study focuses on reducing the sparsity problem through the similarity of items via the values of DC properties, as well as the similarity of users through the values of FOAF properties. The values of properties are heterogeneous type nominal, ordinal, qualitative, etc ., so we have defined several functions of encoding and normalization to convert these properties in a numeric scale. i.e. quantitative values in the range [0-1].

3.1 RDF Specification

Resource Description Framework RDF (http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/) is a data model for the description of various types of resources (person, web page, movie, service, book etc.). It treats the data and its properties and the relationship between them, in other words it is a formal specification by meta-data, originally designed by W3C, whose purpose is to allow a com-munity of users to share the same meta-data for shared resources. However, an RDF document is a set of triplet ¡subject, predicate, object¿ where the subject is the resource to be described, the predicate is the property of this resource and the object it is the value of this property or another resource. One of the great advantages of RDF is its extensibility through the use of RDF schemas that can be integrated and not mutually exclusive with the use of namespace and URI (Uniform Resource Identifier) concepts [7]. It is always possible to present a RDF document by a labelled directed graph. For example, "the book Semantic

Web for the Working Ontologist written by Dean Allemang on July 5, 2011", in RDF/XML Syntax: < ?xml version="1.0"? >

```
<rdf:RDF xmlns:ss="http://workingontologist.org/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<rdf:Description rdf:about="http://www.amazon.fr/
Semantic-Web-Working-Ontologist-Effective/dp/0123859654/">
<ss:written_by rdf:resource="http://www.cs.bu.edu/fac/
allemang/"/> </rdf:Description>
<rdf:Description rdf:about="http://www.amazon.fr/
Semantic-Web-Working-Ontologist-Effective/dp/0123859654/">
<ss:hasTitle>SemanticWeb for the WorkingOntologist</ss:hasTitle>
</rdf:Description>
<rdf:Description rdf:about="http://www.amazon.fr/
Semantic-Web-Working-Ontologist-Effective/dp/0123859654/">
<ss:hasDate >July 5, 2011 </ss:hasDate >
</rdf:Description>
</rdf:RDF>
```

Our solution (figure1) based on a modelling in RDF through FOAF and Dublin core standards,describing the set of the users and items.

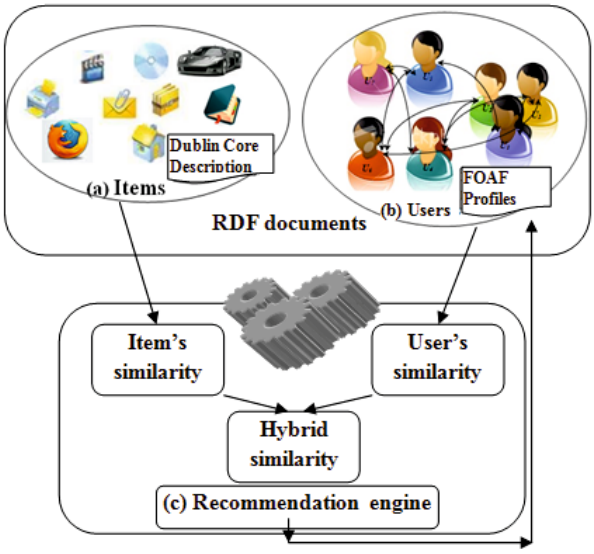


Fig. 1. Overall scheme of the proposal

Thus, in order to keep the collaborative filtering approach we took into account the feedback of the users in the process of computing similarity, moreover we used a hybrid function to define the prediction value. To facilitate the integrity and interoperability, all the documents are represented in RDF/XML notation.

3.2 Item's Representation

A social FS consists of resources items, the users profiles and the histories which memorizes the interactions of the users (ratings) about items recommended. We exploited the meta-data of the Dublin core vocabulary as being a standardization description of items, the attributes values of the vocabulary allowed us to calculate the degree of similarity between items and group them into communities.

Dublin Core vocabulary. Dublin Core DC (*http://dublincore.org*) is a set of simple and effective elements to describe a wide variety of web resources, the standard version of this format includes 15 elements of which semantics has been established by an international consensus coming from various disciplines recommended by W3C. These elements are gathered in three categories those which describe the contents (*Cover, Description, Type, Relation, Source, Subject*) and those which describe the individual properties (*Collaborator, Creator, Editor, Rights*) and others for instantiations (*Date, Format, Identifier, Language*), the current version is known as 1.1, validated in 2007 and revised in 2012 by DCMI (Dublin Core Metadata Initiative, (*http://dublincore.org/documents/dces/*).

Description of items. The core of FS is to form properly the communities, according to well determined criteria, in our research we propose to form the items by taking of account the qualifier DC meta-data QDCMI. We define the set of items as follows:
$I = \{(i_1^1, i_1^2, ...i_1^p), (i_2^1, i_2^2, ...i_2^p), ...(i_m^1, i_m^2, ...i_m^p)\}$ where i_k^j represent the j^{th} property for item k which is identified by its URI and is specified by its qualifiers. We group items by degree of similarity, so I_1 the set of properties assigned to the i_k item and I_2 is the set of properties assigned to the i_l item, then the degree of similarity between i_k and i_l by cosine measurement is given by:

$$sim(i_k, i_l) = \frac{\sum_{j \in I_1 \cap I_2} i_k^j . i_l^j}{\sqrt{\sum_{j \in I_1} \left(i_k^j\right)^2} \cdot \sqrt{\sum_{j \in I_2} \left(i_l^j\right)^2}} \quad (1)$$

This similarity value, allows to group items based on their associated DC properties.

3.3 User's Representation

The objective of FS is to deliver the relevant items to the user, because the formation of the communities depends on the attributes values defined in the user profile. Among the most common current practices we adopted the FOAF vocabulary to represent our profiles.

FOAF vocabulary. FOAF (Friend Of A Friend), is an RDF vocabulary for describing in structured manner a person and his relationships (*http://www.foaf-project.org*). However, it can be used to search for individuals and communities: CV, social networks and management of the online communities, online identification and management of participation in projects etc. A file FOAF can contain various information (*name, family_name, dateOfBirth, gender, mbox, Home Page, weblog, interest, accountName, Knows,etc.*). The major advantage of this representation is the ability to integrate other vocabularies as *DC* (describing a resource), *BIO* (to reveal biographical information), *MeNow* (describing the current status of a person), relationship (to see the type of relation maintained with a person).

Modelling of the user profile. Following the very high number of the users in interaction, it is very important to well form the community as a building block in the FS and assuming one for all and all for one. In order to formulate knowledge, we organized the user profile with categories of FOAF properties and each category c_i associated with a weight w_i, thus we defined the FOAF similarity according to n categories registered in profile by:

$$sim_f = w_1 sim_{c1} + w_2 sim_{c2} + ...w_n sim_{cn} \quad \begin{cases} \sum_i w_i = 1 \\ 0 \leq w_i \leq 1 \end{cases} \tag{2}$$

For our study, we retained three principal categories according to the evolution on the time axis, the first category $c1$, as no evolutionary, includes the non-changeable foaf properties such as: *name, birth_day, gender, mbox,etc.*, the second category in the medium and long term $c2$ contains the foaf changeable properties such as: *account, focus, homepage, phone, skypeID, status, depiction* etc., and the third category $c3$ is defined as category of the preferences includes the foaf properties which interest and preferred by the user like *know, interest, logo, topic_interest, weblog, workplace, based_near, membership* etc. so each class is properly associated with a weight w_i . However, the similarity by foaf properties based on the three categories mentioned above becomes:

$$sim_f = w_1 sim_{c1} + w_2 sim_{c2} + w_3 sim_{c3} \tag{3}$$

Let $u_{f1} = f_1^1, f_1^2, ..., f_1^k$ and $u_{f2} = f_2^1, f_2^2, ..., f_2^k$ the set of the foaf properties of the user u_{f1} and u_{f2} user in a given c_i class, then the value of similarity between these two users by the measurement of cosine that given by the following relation:

$$sim_{ci}(u_{f1}, u_{f2}) = \frac{\sum_{j=1}^{k} f_1^j . f_2^j}{\sqrt{\sum_{j=1}^{k} (f_1^j)^2} . \sqrt{\sum_{j=1}^{k} (f_2^j)^2}} \tag{4}$$

If the value of similarity of two users is close to 1 meant that they belong to the same community.

3.4 Recommendation Engine

The purpose of a FS is to distribute relevant items to users, and avoid a hard task of search in a "big data", the current recommender systems lean on the hybrid approaches which our research is belongs. We have proposed a hybrid similarity based on three types of relationships.

Hybrid similarity. In order to adjust the values of predictions, we conceived a formula to calculate the hybrid similarity, definite as follows:

$$sim_h = \alpha sim_{dc} + \beta sim_f + \gamma sim_r \tag{5}$$

The parameters $\alpha, \beta, \gamma \in [0, 1]$ adjusted by the system administrator according to the efficiency and availability of data.

- sim_{dc}, similarity that using the Dublin Core vocabulary for describing items. By the use of the URI, while identifying item and by exploiting its own metadata allowing reduce the sparsity problem.
- sim_f, similarity which depends on the representation of the profiles by the means of FOAF formalism, in favour of the variety of the fields and the availability of the data in profile, thus, we can overcome the problem of cold start of a new user and to still better forming the communities.
- sim_r, concretize the principal of collaboration through the ratings histories of users to estimate the prediction and to establish the recommendation, so consider their implicit tastes that are often difficult to value by attributes depicted in profile.

Prediction function. Before proceeding to the recommendation task, the system calculates the predicted value of an i item for the active user a, for that, we must select the S most similar items to i_l, then we retain the rating feedback of this user for these S similar items according to the relation:

$$p_{a,l} = \frac{\sum_{m=1}^{s} r_{a,m}.sim_h(i_l, i_m)}{\sum_{m=1}^{s} sim(i_l, i_m)} \tag{6}$$

Where $r_(a, m)$: is the rating value of the current user on the m^{th} similar item. S: size of the most similar items.

Recommendation process . The recommendation process is purely automatic and directly related to the prediction value, so a given item is deemed relevant and deserves to be sent to the user if and only if its predictive value is greater than a given threshold.

$$R_{a,l} = \begin{cases} i_l & recommended\ to\ u_a \quad if\ p_{a,l} \geq \rho \\ i_l & not\ recommended\ to\ u_a \quad otherwise \end{cases} \tag{7}$$

4 Experimentation

This section is devoted to the experimental results of our hybrid solution on real data sets. For evaluation and comparison, we implemented item-CF (item based collaborative filtering) approach widely referenced in Collaborative filtering search [6].

4.1 Datasets

For experimental tests we exploited two sets of data:

- *Book − Crossing* dataset (http://www.informatik.unifreiburg.de/ cziegler/ BX/), a free download dataset for ends of research collected by Cai-Nicolas Zeigler in 2004 from the famous Amazone.com site. The dataset constitutes of 278858 users producing 1149780 votes for 271379 books.
- $foafPub$ dataset (http://ebiquity.umbc.edu/resource/), is a set of data extracted from FOAF files collected during the year 2004, includes 7118 FOAF documents collected from 2044 sites and distributed under the Creative Commons license (v2.0). This set has allowed us to import FOAF properties by SPARQL queries to determine the similarity sim_f.

Our empirical tests require the deployment of a parser to extract FOAF and DC properties through the SPARQL engine of the framework jena 2-6-4 (*https://jena.apache.org/*). Several functions have been defined to aggregate and standardize heterogeneous properties. 80% of the data sets allocated to the training phase and 20% for testing phase.

4.2 Relevance Metrics

To evaluate the method presented in this article, we held a special metric and widely used in the FS, it is MAE, and two other metrics, recall and precision of information retrieval field [16,9].

- MAE: Mean Absolute Error, calculating the mean absolute difference between predictions p_i retained by the system and the real evaluations e_i given by users. This measure is simple to implement and directly interpretable.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - e_i|}{N}$$

- *Precision*: it is the ratio between the number of relevant items returned by the system and the total number of items returned.

$$P = \frac{N_{pr}}{N_r}$$

- *Recall*: it is the ratio between the number of relevant items returned by the system and the total number of existing relevant items in the database.

$$R = \frac{N_{pr}}{N_p}$$

These metrics respectively measures the error, the effectiveness and the quality of FS.

4.3 Results and Discussion

In this section, we discuss the experimental results obtained, for that, we divide the dataset size in two parts, one having a proportion of 80% has dedicated for training phase and the other of proportion of a 20% has dedicated for test phase. From Figure 2, the curves show that the MAE error is minimal in the neigh-

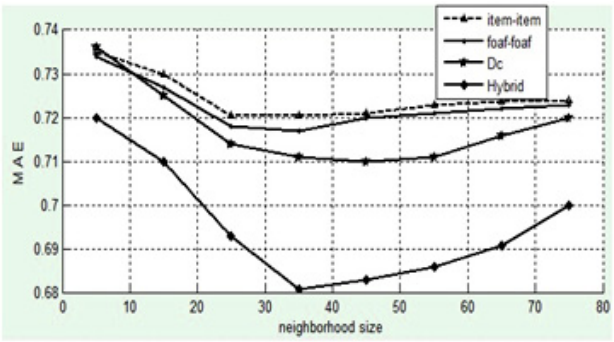


Fig. 2. Comparison of MAE

bourhood range [25-45] and important in outside of this range, it means that as the number of neighbours is less than 25 so there are not enough neighbours to calculate the similarity which lowers the prediction quality, unlike the other side, or the number of neighbours exceeds 45, there are sufficient neighbours, but less similar which degrades prediction quality , this explains that between 25 and 45 there are enough better similar neighbours. Also we observe that the DC curve illustrates a slightly favourable result compared to the FOAF curve, as the items are identified and enriched by descriptions and meta-data with certain stability better than valorising links and subjective opinions between a user′s networks. The best result is obtained in Hybrid curve, or the error is reduced to 0.68 for a neighbourhood size of 35, this favourable result is argued by exploiting items implicit information′s and estimating attributes of user profiles and links between them such as *see also* or *know* properties, which form a social network on the web and therefore a rich database that reduces the MAE, in addition, taking into account the opinions of users through their notes with respect to the items recommended what leads to a profitable collaboration. Two conclusions can be drawn the benefit of this additional data mass reduces the effect of sparsity as a problem moderating filtering systems, and adequately addresses the cold start problem for a new item. Moreover, the URI clause for the unique

resource identification in rdf documents lowers the effect of scalability. In the experiment below, we study the behaviour of our algorithms via the precision and recall metrics. Figure 3 shows a better accuracy rate (up to 73%) for the Hybrid solution, indicates the ability of the system to reject irrelevant items with minimal attribute values.

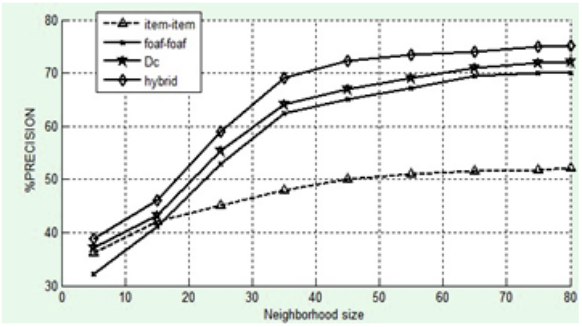


Fig. 3. Precision rate

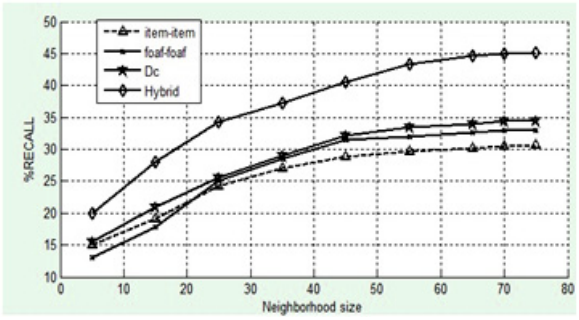


Fig. 4. Recall rate

We also observe that the recall rate (figure 4) which reaches a maximum rate of 45% for the optimal Hybrid solution involves the role of property values of adopted vocabularies to filter only the relevant items.

5 Conclusion and Future Work

Filtering systems are powerful and widely used systems on the web, especially for e-commerce or custom search. Our idea is not to hold closed applications that hide behind a particular data warehouse, but go further, and exploit all

kinds of information and to highlight it for integrity, dissemination and interoperability. In order to alleviate the limitations of collaborative filtering systems, we have presented in this paper, a hybrid model based on the FOAF formalism to better appreciate and enrich user profiles via social networks and information networks. The weighted classification that we have defined for the representation yield more adaptable and flexible profiles and still better adjustable, which alleviate the sparsity problem. On the other hand, the use of DC elements to describe items in a standard way leads to the good development of communities and overcome the problem of cold start for a new resource. The notable progress in the results founded by the formal use of meta-data to describe the valued resources and links with a standard and unified structure. Moreover, the union of similarities adopted for the recommendation is considered a balance between using different data sources and therefore increased the quality of prediction. In our opinion, the system model seems to a network of resources in collaboration with a network of properties describing these resources. The adoption of RDF syntax to the representation and implementation ensures openness, sharing and interoperability of all kind of data on the web, thus allows concretizing and developing semantics via these new practices, we think it is important to study the problem of scalability and reduce the computation time through the reduction techniques of the vector space, thus we also plan to still improve the rate of recall by the semantic disambiguation techniques of the users profiles.

References

1. Sieg, A., Moba, B., Burke, R.: Improving the effectiveness of collaborative recommendation with ontology-based user profiles. In: Proceedings of the 1st Interna Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec 2010 (2010)
2. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Item based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web (WWW 2001), pp. 285–295 (May 2001)
3. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of ACM 35(12), 61–70 (1992)
4. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-theart and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
5. Adomavicius, G., Jingjing, Z.: Stability of Collaborative Filtering Recommendation Algorithms. Citeseer (2012), doi:10.1.1.221.7584
6. Hassanzadeh, H., Keyvanpour, M.R.: Semantic Web Requirements through Web Mining Techniques. International Journal of Computer Theory and Engineering 4(4) (August 2012)
7. Konstan, J.A., Riedl, J., Borchers, A., Herlocker, J.L.: Recommender systems: a GroupLens perspective. In: Recommender Systems, Papers from 1998 Workshop. Technical Report WS98-08. AAAI Press (1998)
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1) (2004)

9. Abrouk, L., Gross-Amblard, D., Cullot, N.: Community Detection In The Collaborative Web. International Journal of Managing Information Technology 2(4) (2010)
10. Albanese, M., dAcierno, A., Moscato, V.F., Persia, A.: A multimedia recommender system. ACM Transactions on Internet Technology (TOIT) 13(1) (2013)
11. Cuong Pham, M., Cao, Y., Klamma, R., Jarke, M.: A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis. Journal of Universal Computer Science 17(4) (2011)
12. Beam, M.A., Michael, A., Kosicki, G.M.: Personalized News Portals: Filtering Systems and Increased News Exposure. Journalism & Mass Communication Quarterly 91(1), 59–77 (2014)
13. Mohammadnezhad, N., Mahdavi, M.: An effective model for improving the quality of recommender systems in mobile e-tourism. International Journal of Computer Science & Information Technology 4(1) (February 2012)
14. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. Proceedings ACM (1994)
15. Bahrehmand, A., Rafeh, R.: Proposing a New Metric for Collaborative Filtering. Journal of Software Engineering and Applications 4, 411–416 (2011)
16. Burke, R.: Hybrid recommender systems: survey and experiments. UserModelling and User-Adapted Interaction 12(4), 331–370 (2002)
17. Meyffret, S., Médini, L., Laforest, F.: Confidence on Collaborative Filtering and Trust-Based Recommendations. In: Huemer, C., Lops, P. (eds.) EC-Web 2013. LNBIP, vol. 152, pp. 162–173. Springer, Heidelberg (2013)

Towards a Recommendation System for the Learner from a Semantic Model of Knowledge in a Collaborative Environment

Chahrazed Mediani[1(✉)], Marie-Hélène Abel[2], and Mahieddine Djoudi[3]

[1] Laboratoire des réseaux et des système distribués, Département d'Informatique, Faculté des sciences, Université Ferhat Abbas de Sétif -1-, Sétif, Algérie
chahrazed_mediani@yahoo.fr
[2] Sorbonne universités, Université de technologie de Compiègne, Compiègne, France
marie-helene.abel@utc.fr
[3] Laboratoire XLIM-SIC et équipe TechNE, UFR Sciences SP2MI, Université de Poitiers, Poitiers, France
mahieddine.djoudi@univ-poitiers.fr

Abstract. Collaboration is a common work between many people which generates the creation of a common task. A computing environment can foster collaboration among peers to exchange and share knowledge or skills for succeeding a common project. Therefore, when users interact among themselves and with an environment, they provide a lot of information. This information is recorded and classified in a model of traces to be used to enhance collaborative learning. In this paper, we propose (1) the refinement of a semantic model of traces with indicators calculated according to Bayes formulas and (2) the exploitation of these indicators to provide recommendations to the learner to reinforce learning points with learners, of his/her community of collaboration, identified as "experts".

Keywords: Collaboration · Trace · Indicator · Recommendation system

1 Introduction

The advent of Information and Communication Technologies and particularly Web 2.0 technologies have facilitated learning based primarily on exchanges and resource sharing between learners of the same community (Abel, 2008). On its side, collaborative learning is a process leading to the progressive construction of knowledge. This learning derived from the current of constructivism allows a person to build knowledge from interaction with his surroundings. When these interactions are performed using digital technology, they leave traces. These traces are usually saved in a model of traces (Settouti et al., 2006) and thus made usable for various purposes such as updating a learner model. Taking account the learner activities within a Computing Environment for Human Learning (CEHL) to guide him in his learning is complex. The learner model allows to consider knowledge of all kinds (preferences, motivations, acquired

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 315–327, 2015.
DOI: 10.1007/978-3-319-19578-0_26

knowledge or not, mistakes, etc.). As part of our work, we focus on the interactions between learners via a CEHL and with a CEHL to make recommendations to guide learner in their learning. To this end, we have chosen to characterize a number of actions that a learner can perform in a CEHL to define learning indicators to establish recommendations. Learning is the result of personal and collaborative actions. We therefore consider the traces resulting of these two axes.

To do this, we have chosen to refine the collaboration model of traces proposed by (Wang et al, 2014) and illustrated in the environment E-MEMORAe 2.0 (Abel, 2009). So we have introduced measures to estimate some parameters, unmeasurable and unobservable by observable indicators describing the state of the learner activities and the progression of his knowledge when interacting within a community of learners.

In the following, we state our problem before presenting the limitations of existing work related to measures established to make recommendations to learners. We then detail our approach based on a model of traces increased by indicators and its exploitation through a case study before concluding and advancing the prospects for this work.

2 Motivation

The Information and Communication Technologies and the emergence of collaborative learning platforms have enabled the implementation of collaborative CEHL and related issues such as the lack of information on the learner evolution within the community and the state of his knowledge and his activities in his group. This information is needed to measure the contribution of each member in the community and may be useful in defining the responsibilities of each member of the group. This information is also useful for the learner himself; this allows him to have a state of his learning and to allow him to prepare himself for a more relevant evaluation. To remedy these problems, the analysis of the leaner interaction traces with a learning environment has become a research topic that is rapidly evolving.

3 Related Work

In the context of CEHL, trace-based study is not just about how to analyze the traces but also how to complete them and exploit them to improve learning (Ollagnier-Belbame et al., 2007). Among the works that have been done in the context of CEHL to support observation, we can mention the work treating the analyzing of the learner behavior and the characterization of his activities (Georgeon et al., 2006), and those that treat the interpretation of learner interactions with computing environments and with other users (Siebra et al., 2005), (George, 2004). There are several learning environments where interactions between the system and users are traced, we mention, for example: the collaborative learning environment Drew (Dialogical Reasoning Educational Web tool) (Corbel and al., 2002). COLAT tool (Collaboration Analysis Tool) (Avouris et al., 2004) is an independent tool for any learning system for the analysis of collaborative activities from the log files and video recordings. Recently, much

work has been done to automate, acquire and distribute knowledge. For example, AdaLearn (Alian Al-Akhras and 2010) is an adaptive learning environment that saves learner responses in his profile to latter allow to direct him through recommendations. (Sani et al., 2012) propose an ontology-based architecture to model the learner and adapt learning styles to learners' profiles. (Li et al., 2012) define an original traces model that distinguishes private actions, individual, collective and collaborative. (Wang et al. 2014) define a method to exploit this model based on TF-IDF method to calculate the index of competence of each learner on a given knowledge. This calculation takes into account the activities of the learner about the knowledge in question, but it does not take into account the acquisition of the knowledge. Under this model, a learner can be proficient in knowledge without being proficient in the knowledge that characterizes it.

4 Our Approach

Our approach is to refine the collaborative model of traces of (Li et al., 2012), and taken up by (Wang et al, 2014), by a number of measures to build indicators on the state of the learner knowledge and the progression of his knowledge within a group in a learning session. Among These parameters: we retain the mastery degree of knowledge represented by a concept. To achieve these goals, we have adopted the following approach: (i) propose a semantic model to measure indicators of the contribution of each student in the group, (ii) estimate the contribution of indicators using Bayesian formulas (Triola 2010), this contribution should take into account the knowledge of the learner and his activities, (iii) propose a set of recommendations to assist the learner in his learning and prepare him for a more appropriate evaluation.

As shown in Figure 1, the architecture of the recommendation system that we propose is composed of two modules operating three models: a trace collection module, a pedagogical content model, a learner model, a collaboration model and a recommendation module. This system is used to collect traces of users and store them in a database of traces. A trace is a time sequence of observed containing all user actions to perform a given task.

The first module of our system treats the collecting primary traces in native format. The second module classifies the primary traces coming from the first module as high-level traces along the trace model (Li, 2013). Depending on the content of traces model and pedagogical content model, algorithms for learning indicators calculations of the learner are applied in the recommendation module. For that, the recommendation system must select good recommendations that guide the user in achieving his learning task. We will illustrate this system of recommendations within the collaborative learning platform E-MEMORAe 2.0 (Abel and Leblanc, 2009).

In the next subsections, we present the principle of the main components of the recommendation system, namely the pedagogical content model, collaborative model, collection of traces and the learner model. The calculation of learning indicators and the recommendation module will be presented in the following sections.

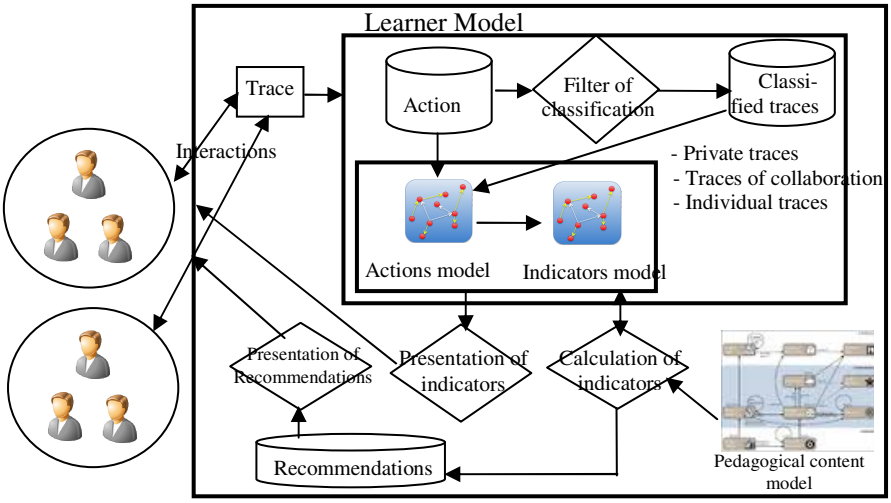


Fig. 1. Architecture of the recommendation system

4.1 Pedagogical Content Model

The content model of learning environments usually consists of a set of elements representing elementary fragments of domain knowledge studied. These elements, often organized in a hierarchy, are named (concepts, notions, knowledge elements, subjects) and they can be of different types. Our content model consists of application ontologies. The application ontology specifies the concepts of a particular application. These concepts represent concepts to be learned of a training unit. A concept is therefore a particular concept that needs to be assimilated by the learner during learning. The concepts are used to index the pedagogical resources treating them. This provides a way to reuse these resources. These concepts are organized in a hierarchy that also represents several types of relationships (specialization and others). Among the ontologies of applications built as part of the E-MEMORAe 2.0 environment: Ontology for the teaching unit "Information Technology". For these applications ontologies, we propose to add the attribute "weight" to the relationship of type "is a" between each concept and its sub concepts (0<= weight<= 1) with the sum of the weights of sub-concepts equal to 1. This value is determined by the responsible of training and represents the degree of contribution of this concept in the acquisition of the father concept (Figure 2).

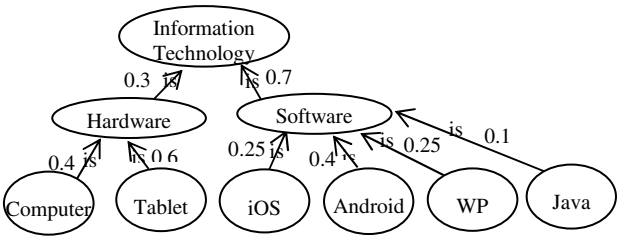


Fig. 2. Part of the application ontology "Information Technology"

4.2 Collaboration Model

Our model allows organizing collaborative spaces for students working in groups on the same problem. Thus forming a work site and exchange for the group and allowing, on the one hand, to each member of the group to access resources (documents and other) for the group and, on the other hand, to memorize his work (documents, ideas, knowledge, solutions, etc.) on the Treaty problem. The MEMORAe 2.0 environment allows each user to choose to access a private space or spaces of groups to which he belongs.

- The private space: space where each user can set his own resources. The content of this space is accessible only by that user.
- Space group: space is only accessible by members of the group and in which they share and exchange resources.

4.3 Collect of Traces

The traces collection is to observe the student in a learning situation and memorize his activities traces to infer the learner model. This collection mode is interesting because it captures the learner interaction the learner without distracting him from his main task. In E-Memorae2.0, these actions are stored in the database traces and classified according to the actions model of the E-MEMORAe 2.0 platform. According to this model, we have three types of traces: Private traces belonging to the private space, traces of collaboration that belong to the space of collaboration and individual traces that are private traces and traces of collaboration. For each type of traces, we have three types of activities that can be conducted by the learner: learning resources consultations (documents), resources creation (conversations, meetings, questions, answers, notes and wikis) and resources additions (documents and annotations).

Example: Figure 3 shows an example of interaction on different concepts, of a group of users using a histogram. Each line represents the collaboration traces of a user for each concept.

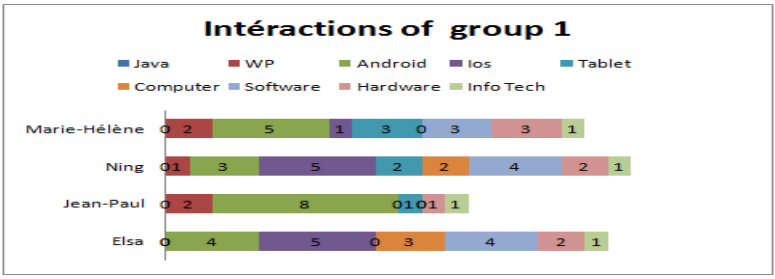


Fig. 3. Example of collaborative interactions in a group

The following table summarizes the actions, of figure 2, performed by members of the group 1. For a given concept, each cell of the table represents the number of actions performed by the learner for each type of activity (C: Consultation, R: Creation,

A: Addition). The number before the parenthesis is the sum of the learner actions for the concept.

Table 1. Summary of users' actions of group 1

	Elsa	Jean-Paul	Ning	Marie-Hélène	Total
Java	0(0C,0R,0A)	0(0C,0R,0A)	0(0C,0R,0A)	0(0C,0R,0A)	**0(0C,0R,0A)**
WP	0(0C,0R,0A)	2(1C,1R,0A)	1(0C,0R,1A)	2(0C,1R,1A)	**5(1C,2R,2A)**
Android	4(1C,3R,0A)	8(4C,3R,1A)	3(1C,1R,1A)	5(4C,1R,0AS)	**20(10C,8R,2A)**
Ios	5(2C,1R,2A)	0(0C,0R,0A)	5(2C,2R,1A)	1(0C,0R,1A)	**11(4C,3R,4A)**
Tablet	0(0C,0R,0A)	1(0C,1R,0A)	2(1C,0R,1A)	3(0C,2R,1A)	**6(1C,3R,2A)**
Computer	3(2C,0R,1A)	0(0C,0R,0A)	2(0C,0R,2A)	0(0C,0R,0A)	**5(2C,0R,3A)**
Software	4(2C,1R,1A)	0(0C,0R,0A)	4(1C,2R,1A)	3(1C,1R,1A)	**11(4C,4R,3A)**
Hardware	2(1C,1R,0A)	1(0C,1R,0A)	2(1C,0R,1A)	3(0C,2R,1A)	**8(2C,4R,2A)**
Info_Tech	1(1C,0R,0A)	1(1C,0R,0A)	1(1C,0R,0A)	1(1C,0R,0A)	**4(4C,0R,0A)**
Total	**19(9C,6R,4A)**	**13(6C,6R,1A)**	**20(7C,5R,8A)**	**18(6C,7R,5A)**	**60(28C,24R,18A)**

C : Consultation, **R** : Creation, **A** : Addition.

4.4 Learner Model

Our learner model is a subset of the pedagogical content model. The pedagogical content is decomposed into a set of elements and the learner model is represented by a set of measurable values associated to these elements. These values vary between 0 (not mastered) and 1 (mastered). The structure of the learner model is the same as the Bayesian network (Figure 4). The elements (concepts and activities) of the learner model become nodes in the Bayesian network. The weight of each element is replaced, for each variable, by a probability to estimate the mastery degree of the learner knowledge. These probabilities vary between 0 (not mastered) and 1 (mastered). The relationship of type "is-a" in the learner model become conditional dependencies between variables forming arcs of the Bayesian network. The elements of knowledge or concepts represent unobservable variables while other elements which are the learning activities used to measure the mastery degree of the learner knowledge (tests, exercises, forums, etc.), represent the observable variables which are added to the Bayesian network.

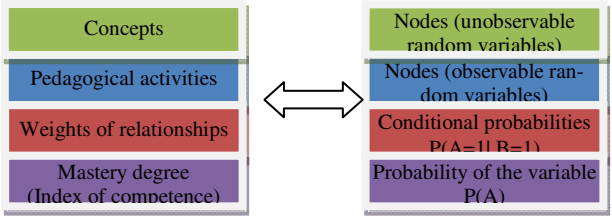


Fig. 4. Relationship between our learner model and a Bayesian network

5 Construction of the Learner Model from Indicators

For a given concept, we aim to measure the degree of mastery of this concept by the learner from the activities related to the concept which he conducted in his group space (contribution by activities) and also from the knowledge acquired by navigating the sub concepts and implementation of activities related to his sub-concepts (contribution by the sub-knowledge). For each concept, we assign a weight P1 to the contribution activities and a weight P2 to the contribution by the sub-concepts. The sum of these weights must be equal to one.

Example: P1 = 0.6, P2 = 0.4.
Each concept is linked to a set of activities, so we can estimate the degree of the learner contribution based on the number of the learner activities carried out in his group. We assign to each type of activity a weight (parameter) which represents the degree of contribution of his activities in the calculation of the mastery degree of this concept by the learner. This setting can distinguish concepts that require a more theoretical activity (consultation) from practical (realization of an exercise, creating a resource). The sum of the weights of the types of activities should be equal to one.

Example: For a given concept, Poids_consultation = 0.2, Poids_création = 0.5 and Poids_addition = 0.3. (These weights can vary from one concept to another).

To realize our contribution model, we use the format of Resource Description Framework (RDF). RDF graph is a model that is used to formally describe Web resources and metadata. Figure 5 shows the RDFS graph of our knowledge model. An ellipse is a class resources and a rectangle represents a property.

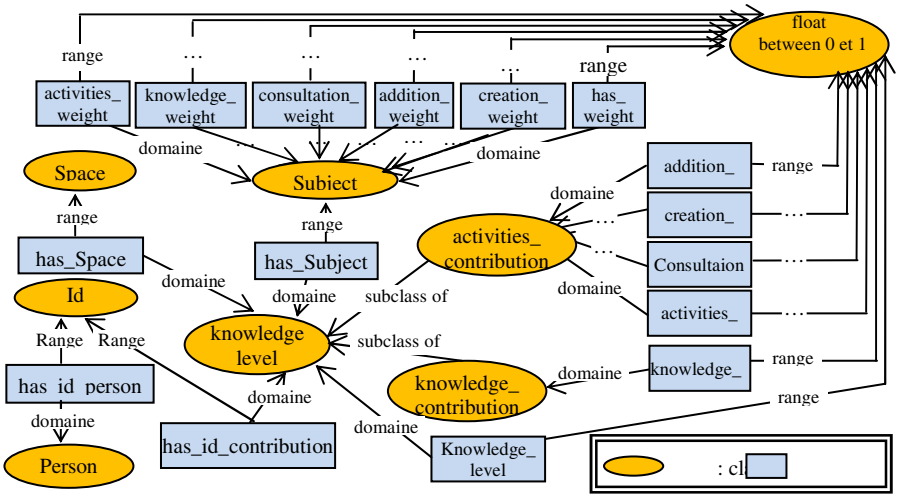


Fig. 5. The knowledge model in the platform E-MEMORAe 2.0

5.1 Indicators Calculation

To measure learning indicators (mastery degree of a concept, degree of contribution by activities, degree of contribution by the sub concepts), of the learner in his space group, we apply Bayesian formula. The Bayesian network is a probabilistic technique that has been developed in the research context to describe the uncertainty of facts in artificial intelligence. Bayesian networks allow easy representation of causal relationships in the learner model. Generally, the learner model information is related to each other. In other words, the learner's knowledge affects each other, for example, motivation to learn, has a direct influence on the ability to perform the task. And as the student model has an hypothetical character, using probabilities, uncertainty is processed. The calculation we retain is based on the following formula:

Consider a partition A_1, A_2, …, A_n of the set E of events: $A_1 \cup A_2 \cup \ldots \cup A_n = E$, $A_i \cap A_j = \varnothing$ for $i \neq j$, $P(E) = 1$. For any event B:

$$P(B) = P(A_1) \cdot P(B \mid A_1) + P(A_2) \cdot P(B \mid A_2) + \ldots + P(A_n) \cdot P(B \mid A_n). \quad (1)$$

P(B | A) : the conditional probability that event B is true given that the event A is already true. We apply equation (1) to calculate the previous indicators.

The Contribution by the Activities

For a learner i in a space S, the index of the contribution by activities AC(i, j) for a concept j is calculated as follows:

$$AC(i, j) = \sum_{k=1}^{n} P(k) * contribution\_value(k) \quad (2)$$

With n: the number of types of activities, in our case n = 3. P(k): the weight of the type of activity k (consultation, creation or addition), contribution_value(k) : is relative frequency estimated by the ratio between the number of activities of type k performed by the learner in the group and the number of all activities of type k performed by all the members of the group S. contribution_value is either consultation_value, creation_value or addition_value.

Example: for the concept “Android", suppose that: Poids_consultation = 0.2, Poids_création = 0.5 and Poids_addition = 0.3. Using table1, we calculate the contribution by the activities of the members of group 1 for this concept (Table 2).

Table 2. Contributions by the activities of the users of group1 for "Android"

	Consultation_value	Creation_value	Addition_value	Activities_value
Elsa	1/10=0.1	3/8 = 0.375	0/2 = 0	0.1*0.2+0.375*0.5+0*0.3 = 0.1975
Jean-Paul	4/10= 0.4	3/8 = 0.375	1/2 = 0.5	0.4*0.2+0.375*0.5+0.5*0.3 = 0.4175
Ning	1/10 = 0.1	1/8 = 0.125	1/2 =0.5	0.1*0.2+0.125*0.5+0.5*0.3 = 0.2325
Marie-Hélène	4/10 = 0.4	1/8 = 0.125	0/2 =0	0.4*0.2+0.125*0.5+0*0.3 = 0.1425

We calculate the contribution by activities of Marie-Hélène for all the concepts.

Table 3. Contributions by the activities of "Marie-Hélène" for all the concepts

	Java	WP	Android	Ios	Tablet	Computer	Software	Hardware	Info_Tech
Marie-Hélène	0	0.4	0.142	0.075	0.483	0	0.274	0.4	0.05

The Contribution by sub-knowledge .

For a learner i, the index of contribution by sub-knowledge for a concept j KC(i, j) is equal to:

$$KC(i, j) = \sum_{k=1}^{n} P(k) * knowledge\_level(k) \quad (3)$$

n is the number k of the sub-concepts related to the father concept j. P (k): the weight attached to each sub-concept k. (These weights are defined in the ontology of application).

Example: The concept "Android" has no sub concepts thus:

$$KC(\text{Marie-Hélène, Android}) = 0$$

The Global Contribution (Mastery Degree)

Now, the mastery degree or the knowledge level of the learner i on the concept j KL(i, j) is equal to:

$$KL(i, j) = P1 * AC(i, j) + P2 * KC(i, j) \quad (4)$$

P1 and P2 are the weights connected to both contributions (activities_contribution and knowledge_contribution respectively).

Example: The concept "Android" has no sub-concept. So the only contribution for this concept is the contribution by activities (P1 = 1, P2 = 0).

KL(Marie-Hélène, Android) = P1 * AC (Marie-Hélène, Android) + P2 * KC (Marie-Hélène, Android)

$KL_{(\text{Marie-Hélène, Android})} = 1*0.1425 + 0*0 = 0.1425$

Table 5 summarizes Marie-Hélène mastery levels for high-level concepts.

Table 4. Knowledge levels of "Marie-Hélène" for the sub-concepts

	Java	**WP**	**Android**	**Ios**	**Tablet**	**Computer**
Marie-Hélène	0	0.4	0.142	0.075	0.483	0

Now, we will infer the mastery degree of Marie-Hélène for the concepts Software, Hardware and Information Technology. Suppose, for these concepts, the weight attached to activities P1 = 0.6 and the weight attached to the sub-concepts P2 = 0.4.

$KL_{(\text{Marie-Hélène, Software})} = P1 * AC_{(\text{Marie-Hélène, Software})} + P2 * KC_{(\text{Marie-Hélène, Software})}$

By applying equation (3):

$KC_{(\text{Marie-Hélène, Software})} = 0.25*0.075 + 0.4*0.142 + 0.25*0.4 + 0.1*0 = 0.159$

Applying equation (4) :

$$KL_{(Marie\text{-}Hélène,Software)}=0.6*0.274+0.4*0.159= 0.228$$

Table 5 summarizes Marie-Hélène mastery levels for high-level concepts.

Table 5. Knowledge levels of "Marie-Hélène" for high-level concepts

	Software	**Hardware**	**Information Technology**
Marie-Hélène	0.228	0.356	0.119

6 The Recommendation Module

The objective is to generate recommendation knowledge to the learner about his learning path from learning indicators stored in the indicators model. A recommendation R consists of an action proposal to achieve.

$$R = <u, s, c, \textit{task}, (o_1,o_2, \ldots,o_n)>$$

- u: the traced user.
- s: the workspace.
- c: the concept concerned by the recommendation.
- task: the task we have to recommend the user to do it. It is either consult resources, add resources, create resources or consult other concepts.
- (o1, o2, ..., on): all users of the space s that can help the user u in achieving the task task.

Recommendation Algorithm

Input: Indicators model, P: Person, S: Space, C: Concept, ε : threshold between 0 and 1/n (n: number of members of the group). RB : Recommandations Base.

Output: Recommendations Knowledge.

```
Indicators := Search_indicators(A,C,S) in the indicators model.
if AC(P, C) < ε then
  if consultation_value < ε then
    U:= search_all_users(S, 'consultation_value> ε')
    Add(<P,S,C,'consult_resources',U>, RB).
  endif
  if addition_value < ε then
    U := search_all_users(S, 'addition_value > ε')
    Add(<P,S,C,'add_resources',U>, RB).
  endif
  if creation_value < ε then
    U := search_all_users(S, 'creation_value > ε')
    Add(<P,S,C,'create_resources',U>, RB).
  endif
endif
if  KC(C) < ε then
```

```
   CO := search_all_sub_concepts(C, 'knowledge_level < ε')
   U := search_all_users(S, 'knowledge_level > ε')
      For all sub-concept Cj of CO do
       Add(<P,S,Cj,'consult_concept',U>, RB).
      endfor
endif
end.
```

According to our model, if the mastery degree of a concept is below a certain threshold. Our algorithm can determine if this is due to the fact that the student has not achieved enough of activities in his workspace. If this is the case, the algorithm also determines if it is consultation, addition or creation activities. And in this case, he recommends the learner to perform more activities and provides a list of students in his group identified as "experts" that can assist him in the implementation of these activities. The algorithm can also determine if the problem is due to the lack of mastery of one or more sub-concepts and in this case, the algorithm recommends the learner to work more on these sub-concepts with learners who have already acquired skills for these sub-concepts.

7 Discussion

Our approach allows measuring some learning indicators such as the mastery level, of a concept by a learner, which is calculated according to the different activities that he has carried out within a group (collaboration space) on the concept and its sub concepts.

Returning to the previous example, we calculated the learning indicators of Marie-Hélène, within the group1 composed of four members, for each concept of the application ontology that contains the concepts to be learned of a training unit. If we apply the recommendation algorithm to Marie-Hélène on the concept “Information Technology” with a threshold of 0.25 (1/4), we obtain:

$KL_{(\text{Marie-Hélène InfTech})} = 0.119 < 0.25$

We have: $AC_{(\text{Marie-Hélène, InfTech})} = 0.05$ (consultation_value = 0.25, creation_value = 0, addition_value = 0). Marie-Hélène will be recommended to work on creating and adding resources with members of his group productive on these actions.

$KC_{(\text{Marie-Hélène InfTech})} = 0.3 * 0{,}356 + 0.7 * 0{,}228 = 0{,}227 < 0.25$

The recommendation also focuses on sub concepts.

$KL_{(\text{Marie-Hélène Software})} = 0.228 < 0.25$ and $KL_{(\text{Marie-Hélène, Hardware})} = 0.356 > 0.25$

The work should be done especially on the sub-concept Software.

Let us now apply the recommendation algorithm to Marie-Hélène on the concept Software with the same threshold, we get:

$AC_{(\text{Marie-Hélène, Software})} = 0.274 > 0.25$ and $KC_{(\text{Marie-Hélène Software})} = 0.159 < 0.25$

Marie-Hélène must therefore be recommended to work on the sub-concepts of Software with learners who have already had expertise on these sub concepts.

We have: $KL_{(\text{Marie-Hélène, Java})} = 0 < 0.25$, $KL_{(\text{Marie-Hélène, WP})} = 0.4 > 0.25$, $KL_{(\text{Marie-Hélène, Android})} = 0.1425 < 0.25$, $KL_{(\text{Marie-Hélène, Ios})} = 0.075 < 0.25$.

These sub-concepts are: Java, Android and Ios.

8 Conclusion

The traces are very important elements in collaborative environments. Their analysis aims to understand and follow the learning of a learner or group of learners and qualify the use, usability and acceptability of collaborative environment to make it more adaptive. In this paper, we proposed an architecture for a recommendation system for the learner. This architecture is based on an original model of the learner taking into account the definition of data (learning indicators). A knowledge base containing this information was constructed. Interaction data recorded were used to construct indicators of learners' state, group state and the progression of the training session. The absence of such indicators in current learning and teaching environments has allowed us to justify our work. We have chosen to use a Bayesian formula to calculate the knowledge level of a learner on a concept of the application ontology describing the pedagogical content of training.

We are currently working to deploy the recommendation module within the environment E-MEMORAe2.0 in order to test it with students from the University of Setif.

References

1. Abel, M.H.: Apport des Mémoires Organisationnelles dans un contexte d'apprentissage. mémoire d'habilitation à diriger des recherches, université de technologie de Compiègne (2008)
2. Abel, M.H., Leblanc, A.: Knowledge Proc of sharing via the E-EMORAe2.0 platform. In: The International Conference on Intellectual Capital, Knowledge Management & Organizational Learning, pp. 10–19 (2009)
3. Avouris, N., Komis, V., Margaritis, M., Fiotakis, G.: An environment for studying collaborative learning Activities. Educational Technology & Society 7(2), 34–41 (2004)
4. Corbel, A., Girardot, J.J., Jaillon, P.: DREW: A Dialogical Reasoning Web tool. In: The International Conference on Information and Communication Techonologies in Education (ICTE), Badajoz, Spain, November 20-23 (2002)
5. George, S.: Analyse automatique de conversations textuelles synchrones d'apprenants pour la détermination de comportements sociaux. Revue Sciences et technologies de l'information et de la communication pour l'éducation et la formation (STICEF) Numéro spécial: technologies et formation à distance, 165-193 (2004)
6. Georgeon, O., Mille, A., Bellet, T.: Abstract: un outil et une méthodologie pour analyser une activité humaine médiée par un artefact technique complexe. Ingénierie des Connaissances IC 2006, Nantes (2006)
7. Li, Q., Abel, M.H., Barthès, J.P.: Facilitating Experience Groups Sharing Collaborative Trace. In: Proceeding of Reuse Exploitation. and In International Conference on Knowledge Management and Information Sharing, pp. 21–30 (2012)
8. Ollagnier-Beldame, M.: A.: Faciliter l'appropriation des EIAH par les apprenants via les traces informatiques d'interactions. Sticef spécial traces (2007)
9. Sani, M.R.F., Mohammadian, N., Hoseini, M.: Ontological learner modeling. Procedia - Social and Behavioral Sciences 46, 5238–5243 (2012)

10. Settouti, L., Prié, Y., Mille, A., Marty, J.-C.: Système à base de traces pour l'apprentissage humain. Colloque international TICE 2006, Technologies de l'Information et de la Communication dans l'Enseignement Supérieur et l'Entreprise (2006)
11. Siebra, S., Salgado, A.C., Brézillon, P., Tedesco, P.: A learning interaction memory using contextual information. The CONTEXT 2005 Workshop on Context and Groupware, Paris, France (2005)
12. Triola, M. F.: Baye's Theorem. Pearson education (2010)
13. Wang, N., Abel, M.H., Barthès, J.P., Negre, E.: Towards a Recommender System from Semantic Traces for Decision Aid. KMIS, Rome (October 2014)

Toward a New Recommender System Based on Multi-criteria Hybrid Information Filtering

Hanane Zitouni[1(✉)], Omar Nouali[2], and Souham Meshoul[1]

[1] Department of Computer Science, University Abdelhamid Mehri, Constantine, Algeria
h_zitouni@esi.dz, smeshoul@gmail.com
[2] Department of Research Computing, CERIST, Algiers, Algeria
onouali@cerist.dz

Abstract. The Communities of Practice of E-learning (CoPEs) are virtual spaces that facilitate learning and acquisition of new knowledge for its members. To achieve these objectives CoPE members exchange and share learning resources that can be (online courses, URLs, articles, theses, etc ...). The growing number of adherents to the CoPE increases the number of learning resources inserted into the memory of this learning space. As consequence, access to relevant learning resource and collaboration between members who have similar needs become even more difficult. Therefore, recommender systems are required to facilitate such tasks. In this paper we propose a personalized recommendation approach dedicated to CoPE that we call Three Dimensions Hybrid Recommender System (3DHRS). The approach is hybrid as it uses collaborative filtering supported by content based filtering to eliminate the problems of cold start and new item. Furthermore, it considers three criteria namely role, interest and evaluation to efficiently solve the new user, and sparsity issues. A prototype of the proposed system has been implemented and evaluated through the use of Moodle platform as it hosts many communities of practice. Very promising results in terms of mean absolute error have been obtained.

Keywords: Information filtering · Multi-criteria · Role · Interest · CoPE · Personalized recommendation

1 Introduction

The Communities of Practice of E-learning (CoPEs) allow their members with different roles namely teachers, tutors, learners…etc, among others to collaborate and share their experiences [1]. So they constitute an environment of sharing of learning resources, problems already solved, learned lessons in practice and any other learning option. When conducting learning activities, the members of the CoPE may need to find members who have a similar profile to share knowledge or experts in the field to get advising. They may also need to get easy access to educational resources related to their field of interest and to be constantly informed about new relevant learning resources. In order to fulfil these needs recommender systems are required (RS).

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 328–339, 2015.
DOI: 10.1007/978-3-319-19578-0_27

Recommender Systems (RS) can be defined [2] as system that allows guiding the user in a personalized way to interesting or useful objects in a large space of possible options. RS are a specific type of information filtering (IF) devoted to present information items (movies, music, books, news, images, web pages, etc ...) that are likely to interest the user. Typically, a RS compares the profile of a user to some reference characteristics, and seeks to predict the "opinion" that he would give. These characteristics may come from either the item itself or from the social environment. The first case refers to content-based filtering while the second case refers to collaborative filtering. When both cases are considered hybrid filtering is achieved [3], [4].

Content Based Filtering (CBF) or cognitive filtering [5], [6], [7] is an important topic in information filtering. It is mainly based on comparing contents of documents (topics) to profiles consisting of themes. Each system user has a profile that describes its own interests. On arrival of a new document, the system compares the representation of the document with the profile to predict user satisfaction on this document. Although CBF is an important technique for information filtering, it suffers from Over-specialization: content-based method provides a limit degree of novelty, since it has to match up the features of profile and items. A totally perfect content-based filtering may suggest nothing "surprising".

Collaborative Filtering (CF) is considered as one of the most successful approaches for building recommender systems. It uses behaviours, activities and known preferences of a group of users to predict and make recommendations of the unknown preferences for other users [8].Typically this technique mainly based on an evaluation criterion is known as Classic Collaborative Filtering (CCF).

Unlike CBF, a CCF approach ignores the form and the content of items. Therefore, does not require any kind of document analysis and complex recommendations could be made. However, CCF raises some issues that should be properly addressed [9], [10] namely:

–*First-Rater problem*: also known as new item problem or clod start item. This problem concerns new items with no ratings. It is impossible for the system to recommend such items to someone because they can't be compared to the other products due to the missing ratings.

–*Sparsity problem*: A similar problem occurs if there is a big amount of products in the system and users don't rate too many products. Thus, it is difficult to find sufficiently correlated users.

–*No preferences*: also known new user problem or cold start user. At the beginning, a new user does not have any preference values; this makes impossible to give any recommendations to him, because he cannot be compared to other users.

–*Cold start problem*: This problem occurs at the beginning of use of the system in critical cases where the system lacks data to make personalized filter of good quality.

In order to reap advantage from both information filtering approaches and to deal with their issues as well, we propose architecture of a three dimensions hybrid recommender system (3DHRS) that includes three layers namely a CF layer, a CBF layer and a user layer. The main contribution consists in fostering the CCF within the CF layer by considering two other dimensions besides evaluation dimension namely role and interest, supported by domain ontology.

Following this introduction, we present in section 2 some related work that propose recommender systems in context of e-learning. In section 3, we describe the proposed approach 3DHRS.In section 4, the developed prototype along with the obtained experimental results are described. Finally, conclusion and future work are given.

2 Related Work

In the e-learning domain, several number of recommender systems have been developed. Such systems play an important educational role. The following table 1 reviews some recent approaches.

Table 1. E-learning recommender systems

Systems	Technique	Object(s) recommended	Dedicated to CoPE	Short description
Alterred Vista system [11].	CF	-Learning resources -People (with similar tastes)	No	Clusters users based on the evaluations of learning resources
RACOFI ([12], [13]).	Hybrid recommendation	-Learning resources	No	Combines two recommendation approaches CF and association rules
QSIA ([14], [15])	CF	-Learning resources	Yes	Used in the context of online communities
CYCLADES [16].	CF	-Learning resources	No	Proposed an environment where users search, access, and evaluate (rate) digital resources
A similar sequencing system [17].	Markov chain model	-Learning paths	No	Calculate transition probabilities of possible learning objects in a sequenced course of study
an evolving e-learning system [18].	Hybrid recommendation	-Learning resources	No	Recommendation takes place both by engaging a Clustering Module and a CF module
ReMashed [19].	Hybrid recommendation	-Services	No	Recommendations based on CF combined with Web2.0 sources

The following observations can be made based on features reported on table 1 and a thorough investigation of most developed RS that we conducted in our study:

–A lot of recommender systems are based on CF only or CF combined with another technique;

–Many of these systems recommend only learning resources while a few others recommend other objects like learning paths, services, people with similar tastes, etc ;

–Most of the systems based on CF create the communities of actors on use only the criterion of evaluation (mono-criterion);

–Almost all the systems proposed are not designed for CoPE.

These facts have motivated our work in proposing a new recommendation approach of users and learning resources, based on content based filtering and multi-criteria collaborative filtering. In the following, a detailed description is given.

3 Proposed Approach

Our study on recommender systems focuses on CoPE which is considered as virtual space for exchanging and sharing: problem solutions; of learning resources; services, etc, by the actors of e-learning during their learning process.

In this paper we propose to use personalized recommendation based on information filtering, in order to guide users to valuable resources, and actors in a wide space of options. Indeed, our recommender system will: Recommend valuable resources that can meet the needs of actors, and recommend also expert members who will validate certain knowledge, do suggestion of members who have a similar profile to improve the collaboration and knowledge exchange between different CoPE actors.

3.1 The Basic Concepts

Following are some basic concepts of our approach to recommendation:

- ***User***: Users are the actors of the CoPE. Each user is characterized by: a role, a field of expertise and interests.
- ***Items:*** are learning resources exchanged and shared among different users.
- ***Evaluation*** is a measure of satisfaction about a specific item it can be:
  -*Explicit*: It's a given user rating on a scale of 1 to 5.
  -*Implicit*: The system induces user satisfaction through his actions.
- ***Profile*** is a description for each user. It contains a static part where personal data about the user (name, surname, age, address,…etc.) are saved and a dynamic part that contains dynamic data that like interests, ratings, interactions etc. On arrival of a new document, the system compares the representation of the document with the profile to predict user satisfaction on this document.
- ***Community:*** It is a set of users gather based on a specific criterion.
- ***Recommendation:*** A list of Top-K elements where the target user will like his majority. These elements can be either users or items.
- ***Similarity:*** The similarity is a numerical value that measures the similarity between items or users based on predefined criteria.
- ***Prediction:*** it is a numeric value that estimates whether the user likes or dislikes the recommended item or the user.
- ***Metadata of learning resources***: is data used to define or describe other data. Metadata is used to describe and index the content of the learning resources.

3.2 General Architecture

The general architecture of our 3DHRS encompasses three main layers, as shown on Figure 1, namely: the layer of Collaborative Filtering (CF), Layer of Content Based Filtering (CBF), and the user layer.

The features of these different layers can be described as follows starting from the lowest layer to the highest one.

- ***Collaborative Filtering layer (CF)***

CF layer is the deepest layer. It is considered as the core of 3DHRS. It consists of two sub-layers, one based mainly on technical Classic Collaborative Filtering and the other sub-layer called Multidimensional/Semantic (M/S). Figure 2 shows the general architecture of this layer.

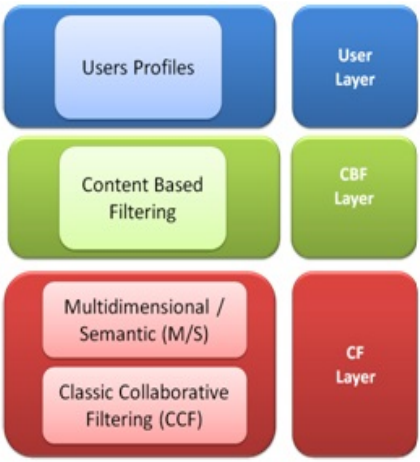


Fig. 1. General Architeture of 3DHRS

✓ *Sub layer Classic Collaborative Filtering (CCF)*

CCF sub-layer relies on a memory based and user centered technique of classic collaborative filtering. It processes the ratings that users have made on certain documents in order to recommend the same documents. Documents should be interesting and of good quality with varied themes. However it requires proper handling of new user and sparsity problems. In our work, we propose a solution to these problems by adding the multidimensional/Semantic (M /S) sub-layer.

✓ *Sub layer Multidimensional / Semantic (M/S)*

According to our study we find that CoPEs members, who have a common role and/or common interest(s), are very often interested by the same resources. Therefore, we suggest to add two dimensions role and interest to the evaluation dimension to foster the CCF sub-layer and eliminate or even minimize the problems of *new user*, and

sparsity. User' s interests are handled using a domain ontology, (in order to discover semantically similar interests with different syntax).

- ***Content Based Filtering layer (CBF)***

At boot a recommender system based on CF suffers from the problem of *cold start*. In fact, the system has no information on users and items. Collaborative filtering methods cannot operate on an empty matrix of ratings. Another instance of this problem is when a *new item* is added and no pre-rating on this item is provided. This causes the system to ignore the item and as a consequence, the item cannot be recommended. The solution that we propose in this context is to use a layer of content-based filtering (CBF) which allows the system to propose items that are close to the best profile by comparing the content of the analyzed resource to interests of users. However, this technique is much more used on text-based resources (where content analysis is not expensive), while in the field of CoPE, resources are of various types (text, multimedia, PDF files, etc). As a consequence, we propose to use metadata [20] describing learning resources; we support the view of [21] and [22] who proposed to support the standard metadata describing the domain ontology. Figure 3 shows the basic principle of this layer.

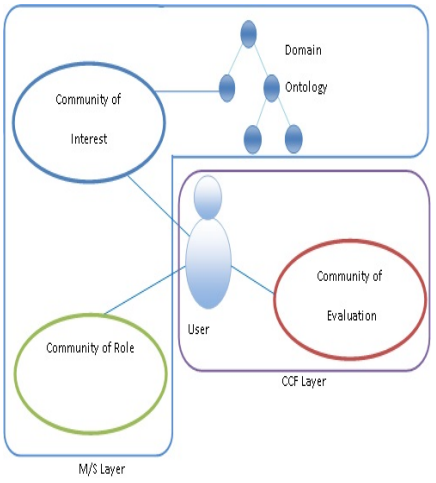


Fig. 2. Collaborative Filtering layer

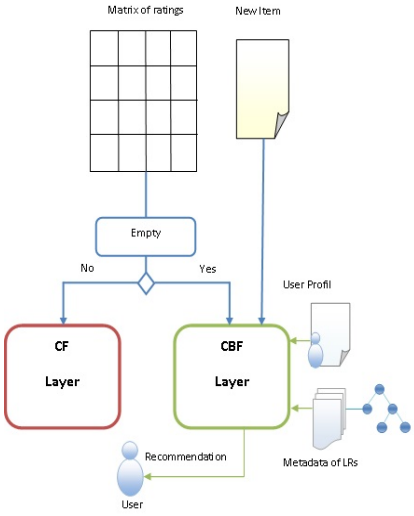


Fig. 3. Content Based Filtering layer

- ***User layer***

This is the only explicit layer for users; his main role is to create user profiles based on the collected data. Figure 4 presents the basic principle of this layer.

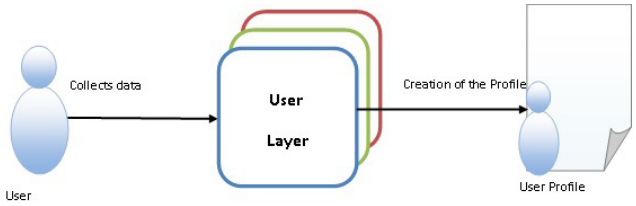


Fig. 3. User layer

3.3 Recommendation Engine

The recommendation engine allows 3DHRS to recommend: *Learning Resources* and *users*. There are four main engine processes of 3DHRS designed to perform the following tasks: pre-evaluation, evaluation of recommendations, creation of communities and finally production of recommendations.

1) Pre- evaluation

Given an empty matrix of ratings or a recently new added item, it is essential to go through a pre-evaluation step. It is mainly based on the CBF.

The basic principle of this step is to compare the interests of the user extracted from his profile to key words describing the items represented in metadata. The comparison between the user preferences and keywords that describe the item is done using the similarity calculation. For this, we adopted the formula of *Jaccard coefficient* defined as:

$$\mathrm{Sim}(i,u)=\frac{|K_i \cap I_u|}{|K_i \cup I_u|} \quad (1)$$

Where: *Sim(i,u)* is a measure of similarity between the user u and the item i, u is the target user, K_i isa set of key-words that describe an item i, I_u is a set of interests of the user u.

2) Evaluation of recommendations

The next step is the evaluation, which provides the ability for users to know the existing items and evaluate them in an explicit way (which is to give a rating on a scale of 1 to 5) or implicitly where the system induces user satisfaction through his actions.

3) Formation of communities

In 3DHRS, communities are formed based on three criteria: roles, interests and user's evaluations.

✓ *Communities of Role/Interest*

The communities of role/interest are communities formed by users who have the same role and/or the same interest based on the matrix of roles and vector of interests.

The matrix of roles is a binary matrix. In order to fill the matrix of roles, it is necessary to calculate the similarity of role based on the following formula:

$$\mathrm{Sim_R(u,x)}=\begin{cases} 0 \text{ if } u,x \text{ have not the same role} \\ 1 \text{ if } u,x \text{ have the same role} \end{cases} \quad (2)$$

Where: u is Target user, x is Any user, $Sim_R(u,x)$ is the measure of similarity of roles between u and x.

To know if two users have one or more common interests, it is necessary to calculate a similarity of interests between them based on their vectors of interests. To measure the degree of similarity of interests between different users, we suggest using the following formula:

$$\mathrm{Sim_I(u,x)}=\frac{|I_u \cap I_x|}{|I_u \cup I_x|} \quad (3)$$

Where: u is Target user, x is Any user, $Sim(u,x)$ is the measure of interest similarity between u and x, Iu is the vector of interests of the user u, Ix is the vector of interests of the user x.

To assign two users to the same community of role/ interest, we measure the degree of similarity of role / interest ($Sim_{R/I}$), which is calculated by the following formula:

$$\mathrm{Sim_{R/I}(u,x)}=\frac{\mathrm{Sim_R(u,x)+Sim_I(u,x)}}{2} \quad (4)$$

✓ *Community of evaluation*

The creation of evaluation community is based mainly on similarity of evaluation. There are three main methods to calculate this similarity: *cosine similarity*, *the modified cosine similarity* and *Pearson correlation coefficient similarity (PCC)*. Many experiments show that the last one can represent the similarity of users or items better than the other methods ([23], [24], [25]). So, we adopted it in order to create the community of evaluation, the formula of PCC is defined below:

$$\mathrm{Sim_E(u,x)}=\frac{\sum_{i\in I_{ux}}\left(R_{u,i}-\bar{R}_u\right)\left(R_{x,i}-\bar{R}_x\right)}{\sqrt{\sum_{i\in I_{ux}}\left(R_{u,i}-\bar{R}_u\right)^2}\sqrt{\sum_{i\in Iux}\left(R_{x,i}-\bar{R}_x\right)^2}} \quad (5)$$

Where: u is Target user, x is Any user, $Sim_E(u,x)$ is the measure of evaluation similarity between u and x, $I_{ux}(I_{ux} = I(u) \cap I(x))$ is the set of evaluated item by u and x, $R_{u,i}, R_{x,i}$ represent evaluations of the user u and x for the item I, $\bar{R}_u, \bar{R}_x$ are respectively the average evaluations from user u and x for all items.

4) Production of recommendations

To produce recommendations of items and users, it is necessary to calculate a prediction of users and items. For this, we propose the following formulas:

$$\mathrm{Pred_U(u,x)} = \frac{\beta_1 \mathrm{Sim_R(u,x)} + \beta_2 \mathrm{Sim_I(u,x)} + \beta_3 \mathrm{Sim_E(u,x)}}{\beta_1+\beta_2+\beta_3} \quad (\mathbf{6})$$

Where: u is target user, x is any user, $\beta_1, \beta_2, \beta_3$ are coefficients where $\beta_1 = \beta_2 = \beta_3 = 1$ except if $Sim_R(u,x) = 0$, $Sim_I(u,x) \neq 0$, $Sim_E(u,x) \geq 0$ and the role of x is « *expert* » in this case β_1=0, $\beta_2 = \beta_3 = 3$.

The main objective behind using coefficients $\beta_1, \beta_2, \beta_3$ is to promote the recommendation of users who have role as "*expert*" and share at least one common interest with the target user.

$$\mathrm{Pred_I(u,i)} = \frac{\alpha_1(\overline{\mathrm{R}}_{\mathrm{R/I,i}}) + \alpha_2(\overline{\mathrm{R}}_{\mathrm{E,i}})}{\alpha_1+\alpha_2} \quad (\mathbf{7})$$

Where: u is the target user, i is an item, α_1, α_2 are coefficients where :$\alpha_1 = \alpha_2 = 1$, R/I is the community of Role/Interest, E is the community of evaluation , $\overline{R}_{R,i}, \overline{R}_{E,i}$ are respectively the average evaluations of role community and evaluation community: with respect to the item i where: $(R/I) \cap E = \emptyset$.

If u is a new user, $\mathrm{Sim_E(u, x)} = 0$ .So, he will not be assigned to a community of evaluation but he may benefit from the recommendations coming from his community of role/ interest This is what we call *initial recommendation*.

After calculating the predictions, we can make a recommendation of a list of Top-K users, and Top-K most predicted items.

4 Implementation and Experimentation

To test the proposed approach it was necessary to find a CoPE. For that we propose to use the plat form moodle[1] that hosts many communities of practice involved in the development of the platform. Among these communities we have: "Moodle Exchange" (ME) offers a virtual place where we can share learning resources in a free community perspective. The figure 5 represents a screenshot of the home interface of Moodle Exchange. Just after the creation of the CoPE: ME we will enhance their environment by the integration of 3DHRS that will provide actors of CoPE: ME a recommendation of resources and users. Figure 6 presents screenshot of this integration. In order to test our prototype, we used as performance measure the *Mean Absolute Error* (*MAE*) which is computed by the following formula:

$$\mathrm{MAE} = \frac{\sum_{u,i} \left| p_{u,i} - n_{u,i} \right|}{n} \quad (\mathbf{8})$$

[1] https://moodle.org/

Where: $n_{u,i}$ is the score given by the user u on item i, $p_{u,i}$ Predicted note, n is the total number of predicted scores.

Fig. 5. Home Interface of Moodle Exchange

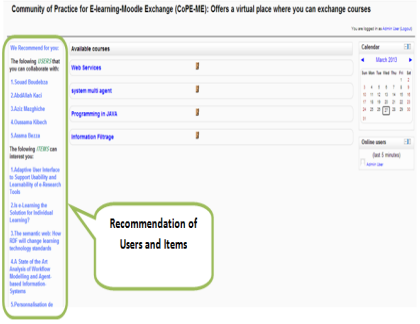


Fig. 6. The Integration of 3DHRS in CoPE: ME

The obtained results are described in Figures 7 and 8 where *K* refers to the number of recommended users and items respectively.

As can be observed on these plots, the values of MAE lie within the range [0.04, 0.37] in the case of users recommendation and [0.06, 0.36] in the case of items recommendations. It is clear that the achieved values are very low which indicate that good quality recommendations have been provided.

In order to show the advantage of the proposed 3DHRS over a CCF approach, a comparative study has been performed. Figures 9 and 10 show the achieved MAE values using both approaches for a list of top-k users (respectively top-k items), where k = 10. We can see that 3DHRS outperforms CCF in case of recommendation of users where 3DHRS MAE values are smaller than those of CCF. In case of recommendation of items, competitive results have been obtained.

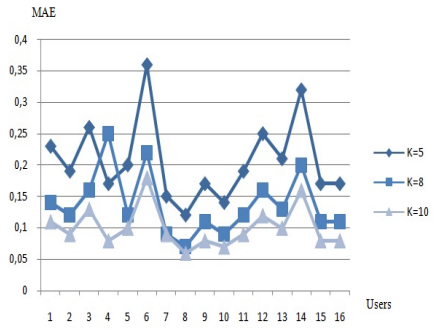


Fig. 7. MAE of Recommendation of top-k Users

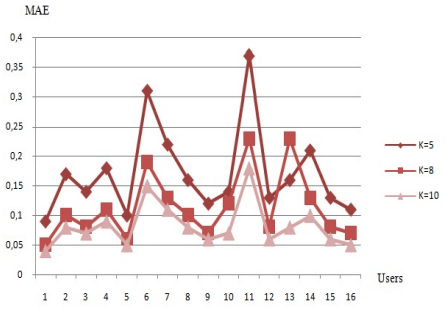


Fig. 8. MAE of Recommendation of top-k Items

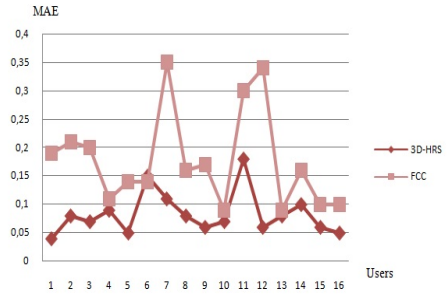


Fig. 4. MAE of 3DHRS VS MAE of CCF in case of users recommendation

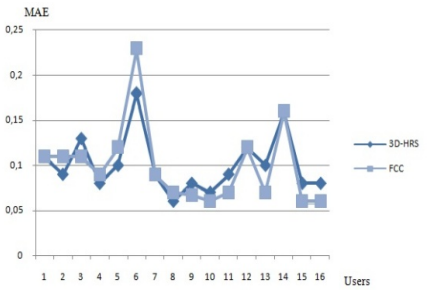


Fig. 10. MAE of 3DHRS VS MAE of CCF . in case of items recommendation.

5 Conclusion

In this paper, we described a new approach to personalized recommendation dedicated for the CoPE, which is mainly based on collaborative filtering supported by the notion of multi-criteria, and combined with content based filtering. Actually we were faced with a challenge to use the technique of information filtering while reducing the impact of the related cold start issues. The proposed approach was implemented using a prototype on which we applied some experiments, the results were very promising.

As future work, it would be interesting to further improve the recommendation of users by adding other mechanisms such as RDF vocabulary (Resource description Framework) and activity concept.

References

1. Hamburg, I.: eLearning 2.0 and Social, Practice-oriented Communities to Improve Knowledge in Companies. In: Fifth International Conference on Internet and Web Applications and Services 2010 (2010)
2. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002); ISSN 0924-1868
3. Ansari, A., Essegaier, S., Kohli, R.: Internet recommendation systems. Journal of Marketing Research 37, 363–375 (2000)
4. Shahabi, C., Banaei-Kashani, F., Chen, Y.-S., McLeod, D.: Yoda: An Accurate and Scalable Web-Based Recommendation System. In: Batini, C., Giunchiglia, F., Giorgini, P., Mecella, M. (eds.) CoopIS 2001. LNCS, vol. 2172, pp. 418–432. Springer, Heidelberg (2001)
5. Lang, K.: NewsWeeder: Learning to Filter Netnews. In: Proceedings of the 12th International Conference on Machine Learning (ICML1995), CA, USA, pp. 331–339 (1995)
6. Lieberman, H.: Letizia: An agent that assists web browsing. In: Proceedings of the14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Canada, pp. 924–929 (1995)
7. Pazzani, M., Billsus, D.: Learning and Revising User Profiles: The Identification ofInteresting Web Sites, Machine Learning, vol. 27, pp. 313–331. Kluwer Academic Publisher (1997)

8. Su, X., Khoshgoftaar, M.T.: A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence, Article ID 421425, 19 pages (2009)
9. Melville, P., Mooney, R.J., Nagarajan, R.: Content-Boosted Collaborative Filtering for Improved Recommendations. In: Proceedings of the 18th National Conference on Arti. (2002)
10. Meier. Community Building Processes Using Collaborative Filtering Information. Thesis on System Research Group University of Fribourg (2008)
11. Recker, M.M., Wiley, D.A.: An interface for collaborative filtering of educational resources. In: Proc. of the 2000 International Conference on Artificial Intelligence, Las Vegas, USA, pp. 26–29 (2000)
12. Anderson, M., Ball, M., Boley, H., Greene, S., Howse, N., Lemire, D., McGrath, S.: RACOFI: A Rule-Applying Collaborative Filtering System. In: Paper presented at the conference IEEE/WIC COLA 2003, Halifax, Canada (October 2003)
13. Lemire, D., Boley, H., McGrath, S., Ball, M.: Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation. International Journal of Interactive Technology and Smart Education 2(3) (2005)
14. Rafaeli, S., Barak, M., Dan-Gur, Y., Toch, E.: QSIA a web-based environment for learning, assessing and knowledge sharing in communities. Computers & Education 43(3), 273–289 (2004)
15. Rafaeli, S., Dan-Gur, Y., Barak, M.: Social Recommender Systems: Recommendations in Support of E-Learning. International Journal of Distance Education Technologies 3(2), 29–45 (2005)
16. Avancini, H., Straccia, U.: User recommendation for collaborative and personalised digital archives. International Journal of Web Based Communities 1(2), 163–175 (2005)
17. Huang, Y.M., Huang, T.C., Wang, K.T., Hwang, W.Y.: A Markov-based Recommendation Model for Exploring the Transfer of Learning on the Web. Educational Technology & So-ciety 12(2), 144–162 (2009)
18. Tang, T.Y., McCalla, G.I.: Smart Recommendation for an Evolving E-Learning System: Architecture and Experiment. International Journal on E-Learning 4(1), 105–129 (2005)
19. Drachsler, H., Pecceu, D., Arts, T., Hutten, E., Rutledge, L., van Rosmalen, P., Hummel, H., Koper, R.: ReMashed - Recommendations for Mash-Up Personal Learning Environments. In: Cress, U., Dimitrova, V., Specht, M. (eds.) EC-TEL 2009. LNCS, vol. 5794, pp. 788–793. Springer, Heidelberg (2009)
20. Zitouni, H., Berkani, L., Nouali, O.: Recommendation of Learning Resources and Users Using an Aggregation-Based Approach. In: Publié au 2ième IEEE Workshop sur les Systèmes d'informations Avancés Pour les Enterprises (IWAISE 2012), Algérie (2012)
21. Bouzeghoub, A., Defude, B., Duitama, J.-F., Lecocq, C.: Un modèle de description sémantique de ressources pédagogiques basé sur une ontologie de domaine, vol. 12 (2005)
22. Abdelwahed, E.H., Lazrek, A.: Des ontologies pour la description des ressources pédagogiques et des profils des apprenants dans l'elearning (2006)
23. Breese, J., Hecherman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, UAI 1998, pp. 43–52 (1998)
24. Billsus, D., Pazzani, M.J.: Learning Collaborative Information Filters. In: Proceedings of ICML 1998, pp. 46–53 (1998)
25. Jun Feng, Z., Xian, T., Jing Feng, G.: An Optimized Collaborative Filtering Recommendation Algorithm. Journal of Computer Research and Development 14(10), 1842–1847 (2004)

Information Technology: Ontologies

A New Approach for Combining the Similarity Values in Ontology Alignment

Moussa Benaissa(✉) and Abderrahmane Khiat

LITIO Laboratory, University of Oran1 Ahmed Ben Bella,
B.P 1524 El M'Naouar, 31000, Oran, Algeria
moussabenaissa@yahoo.fr,
abderrahmane_khiat@yahoo.com

Abstract. Ontology Alignment is the process of identifying semantic correspondences between their entities. It is proposed to enable semantic interoperability between various knowledge sources that are distributed and heterogeneous. Most existing ontology alignment systems are based on the calculation of similarities and often proceed by their combination. The work presented in this paper consists of an approach denoted PBW (Precision Based Weighting) which estimates the weights to assign to matchers for aggregation. This approach proposes to measure the confidence accorded to a matcher by estimating its precision. The experimental study that we have carried out has been conducted on the Conference[1] track of the evaluation campaign OAEI[2] 2012. We have compared our approach with two methods considered as the most performed in recent years, namely those based on the concepts harmony and local confidence trust respectively. The results show the good performance of our approach. Indeed, it is better in terms of precision, than existing methods with which it has been compared.

Keywords: Ontologies · Ontology alignment · Ontology matching · Semantic correspondences · Similarity · Aggregation of the similarities · Combination of the similarities

1 Introduction

The Semantic Web Community, defined as a futuristic extension of the current web, has adopted ontologies as the cornerstone for its achieving in order to overcome the crucial problem of semantic heterogeneity that is inherent to its distributed and open nature. However, these ontologies are themselves heterogeneous. This heterogeneity may occur at syntactic, terminological, conceptual or semiotic levels [5].

Ontology alignment, defined as the process of identification of semantic correspondences between entities of different ontologies to be aligned [5], is proposed

[1] http://oaei.ontologymatching.org/2012/conference

[2] OAEI (Ontology Alignment Evaluation Initiative) organizes evaluation campaigns aiming at evaluating ontology matching technologies. http://oaei.ontologymatching.org/

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 343–354, 2015.
DOI: 10.1007/978-3-319-19578-0_28

as a solution to the problem of semantic heterogeneity by enabling the semantic interoperability between various sources of information.

We globally distinguish two approaches to identify the alignment between ontologies: reasoning-based approaches and those based on the calculation of similarities [12].

Most of the existing ontology alignment systems are based on the calculation of similarities between entities to align. In this category, we distinguish two types of systems: (1) systems which implement one single technique and (2) systems which combine several techniques, in order to estimate the similarity between two entities. The latter systems have become more frequent due to their flexibility and their easy extension [7]. Moreover, with the increasing complexity of ontologies on the Web (number and volume), the alignment cannot be performed reasonably in a purely manually way. Therefore it is imperative to develop automatic or at least semi-automatic systems to identify the alignment [11]. This situation is dictated by the lack of human expert especially in dynamic systems and by the concern to accelerate the alignment process [1].

Precisely, we propose in this paper an ontology alignment approach based on the calculation of similarities and which fits into the category of methods that combine several matchers. It is a statistical approach based on two heuristics to aggregate similarity values calculated by different matchers. The first estimates the candidate final alignment from the alignments identified by matchers, considering their intersection. The second provides an estimate of the weight to be assigned to the matchers with a view of their combination using a weighted summation strategy.

The rest of the paper is organized as follows. In the Section 2, we present some preliminary notions on ontology alignment in order to facilitate the reading of the paper content. The Section 3 contains the description of some related work to our approach. In the Section 4, we present an example in order to illustrate our approach. The Section 5 is dedicated to the presentation of the proposed approach. The Section 6 contains the experimental results obtained during the evaluation of our approach. Finally we give a conclusion and some future perspectives.

2 Preliminaries

In this section we present some preliminary notions of ontology alignment in order to facilitate the reading of the paper content. We outline the notions of ontology, similarity calculation techniques and alignment, respectively. We refer the reader, for more details, to the following references [5] [4].

2.1 Notion of Ontology

Definition: Ontology is a six tuple [2]: $O = <C, R, I, H^C, H^R, X>$ where:

- C: set of concepts.
- R: set of relations.
- I: set of instances of C and R.
- H^C: denotes a partial order relation on C, called hierarchy or taxonomy of concepts. It associates to each concept its super or sub-concepts.

- H^R: denotes a partial order relation on R, called hierarchy or taxonomy of relations. It associates to each relation its super or sub-relations.
- X: set of axioms.

2.2 Techniques of the Similarities Calculation

There are basically five types of methods to calculate similarities [1]:

1. *Terminological Methods*. These methods are based on string matching and can be applied to the names, labels and descriptions of the entities. We cite as an example of matcher of this category: the edit distance.
2. *Linguistic Methods*. These methods are based on external resources as dictionary and thesaurus in order to calculate the similarities between the names, labels and descriptions of the entities. We cite as an example of a matcher of this category: similarity based on WordNet (Wu-Palmer).
3. *Structure-based Methods*. These methods exploit the internal structure (domain, range, properties and cardinality, etc.) and the external structure (hierarchy and the relation-ship between other entities) of the entities in order to calculate their similarities. We cite as an example of a matcher of this category: Resnik similarity.
4. *Semantic-based Methods*. These methods are essentially deductive and inferential and are based on formal semantic of generic or specific domains. We cite as an example of a matcher of this category: SAT solvers.
5. *Instance-based Methods*. These methods exploit the instances associated to the concepts (extensions) to calculate the similarities between them. We cite as an example of a matcher of this category: Jaccard similarity.

2.3 Notion of Ontology Alignment

The alignment of two ontologies is the process of identification of semantic correspondences between their entities. In this section, we briefly introduce the basic necessary concepts on the alignment in order to facilitate the reading of the paper content.

2.3.1. Notion of Correspondence

Let O and O' two ontologies. A Correspondence M between O and O' is quintuple < Id, e, é, r, n > where:

- Id: is a unique identifier of the correspondence M;
- e and e‘ are the entities of O and O' respectively (concepts, relations or instances);
- r: is the semantic relation between e and e‘ (equivalence (≡), more specific (⊆),more general (⊇), disjunction (⊥));
- n: is a measure of confidence, typically a value within [0, 1].

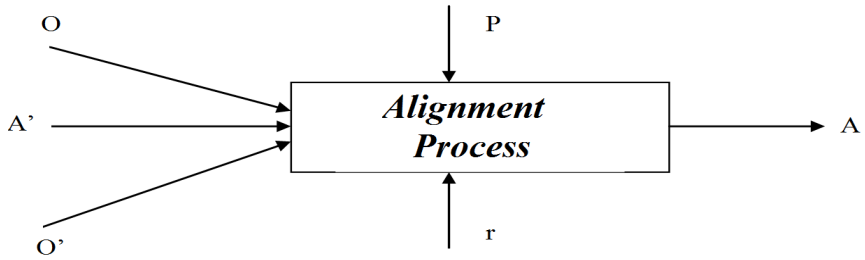


Fig. 1. Alignment Process

2.3.2. Notion of Alignment

The alignment can be defined as a set of correspondences. The alignment process (Fig. 1) receives as input two ontologies O and O' and produces as output an alignment A between entities of O and O'. Other elements complete this definition, namely:

- An initial alignment A' to be completed or refined by the process.
- The external resources r such as a thesaurus or a dictionary.
- The parameters P such as thresholds or weights.

The alignment process consists generally of the following steps:

1. *Analysis*: This step consists of extracting both the entities (concepts, relations, instances) of the two ontologies O and O ' and their characteristics which will be used to identify the alignment.
2. *Calculation of Similarities*: this step consists to execute the different matchers in order to calculate the similarities between entities to align.
3. *Similarity Values Aggregation*: This step consists to combine the similarity values calculated by the matchers in the previous step, into one value.
4. *Selection*: This step consists of applying a strategy, for example a threshold strategy in order to filter the alignment defined in the previous step. Other optimization techniques can also be applied at this level to optimize the extraction of the final alignment.
5. *Improvement of the Alignment*: descriptive logic techniques can be applied at this level to improve the final alignment by diagnosing and repairing any inconsistencies identified in the final alignment.

3 Related Works

The aggregation of similarity values calculated by different matchers consists to combine them into one single value. There are basically three types of approaches to achieve this aggregation: the **weighting**, the **vote** and the **argumentation** [5]. In the vote strategy, the matchers are considered as independent sources of information and the decision to include a correspondence in the alignment is taken on the basis of a

simple majority vote by the matchers for this correspondence. The argument strategy allows negotiating an alignment by exchanging arguments between agents. In the weighting strategy several techniques are proposed to combine the similarity values.

In [3], the authors quote the following strategies to combine similarity values calculated by different matchers: (1) Max: this strategy selects the maximum similarity value among the values calculated by different matchers); (2) Min: this strategy selects the minimum similarity value among the values calculated by different matchers); (3) Average (this strategy calculates the average value of the similarities calculated by different matchers); and (4) Weighting (this strategy calculates the weighted sum of the similarities calculated by different matchers). The latter, which is more frequent in ontology alignment systems [7], requires an estimate of the weights that reflect the importance of each matcher. In some systems this weights approximation is done manually by a human expert. This approach is difficult to implement given the enormous number of possible configurations [11] and has the major drawback to run correctly on a specific alignment task and not on another. It is therefore suitable that the weights estimation be specific to the current alignment task [8].

Several studies have addressed the problem of the weights estimation of different matchers. In [14], the authors propose an approach based on information theory and estimate the weight of each matcher based on the calculation of entropy (uncertainty of information) from the similarity values calculated by this matcher.

The works described in [9] and [13] present an approach based on genetic algorithms to give an estimate of weights assigned to different strategies used.

In [8] the authors propose the harmony concept for weighting the different matchers. The harmony **h** of a similarity matrix **sim** of **n** rows and **m** columns is defined by: « *the number of pairs of entities (e_i, e'_j) for which the similarity **sim** (e_i, e'_j) is the maximum at the same time on the row i and column j, divided by the maximum number of concepts of ontologies to align O and O'* ". This value **h** is assigned as weight to the matcher associated to the matrix **sim**. In [2] the authors propose a local confidence measure for a pair of entities unlike that proposed in [8], which is global to the entire similarity values matrix. This measure, denoted m, is defined for an entity e of the ontology O, by: $m = m_r - m_{nr}$ where m_r is the average of similarity values of entities that are associated to e and m_{nr} is the average of similarity values of entities that are not associated to e.

Other works such as [6] and [10] use machine learning techniques for automatic configuration of weights to be assigned to the matchers.

* The approach proposed in this paper is situated in the category of weighting techniques that combine the similarity values calculated by different matchers. It consists of a heuristic that estimates the weights to assign to the matchers. Contrary to the techniques mentioned above, this approach is of statistical nature and estimates the weights by an estimation of the precision standard metric.

4 Illustrative Example of the Approach

Let two ontologies O and O' which contain the concepts O: {Product, Provider, Creator} and O': {Book, Translator, Publisher, Writer} respectively.

The application of the edit distance metric and that based on WordNet between concepts of O and O' has generated the following two matrices of similarities.

1) If we filter out the matrix of similarities (Table 1) calculated with the edit distance, with a threshold s = 0.15 we obtain the following alignment:
A1= {(Product, Translator), (Provider, Translator), (Provider, Publisher), (Provider, Writer), (Creator, Translator), (Creator, Writer)}.

Table 1. The Similarity Values Calculated by Edit Distance

O /O'	Book	Translator	Publisher	Writer
Product	0.14	0.20	0.11	0.14
Provider	0.12	0.20	0.44	0.50
Creator	0.14	0.50	0.11	0.43

2) If we filter out the second matrix of similarities (Table 2) calculated using WordNet, with a threshold s = 0.15 we obtain the following alignment:
A2= {(Product, Book), (Product, Writer), (Provider, Writer), (Provider, Book), (Creator, Book), (creator, Translator), (Creator, Writer)}.

Table 2. The Similarity Values Calculated Using WordNet

O /O'	Book	Translator	Publisher	Writer
Product	0.18	0.12	0.12	0.15
Provider	0.17	0.11	0.14	0.29
Creator	0.18	0.47	0.12	0.15

The alignment A which consists of the semantic correspondences identified by the two matchers simultaneously is as follows: A = A1 ∩ A2 = {(Creator, Translator), (Provider, Writer), (Creator, Writer)}. A represents the estimator of final candidate alignment.

The estimator of the precision of the matcher edit distance is: P_1=3/6=0.50.

The estimator of the precision of the matcher based on WordNet is: P_2=3/7=0.43.

The weights to be assigned to matchers are: w_1 = 0.50 and w_2 = 0.43.

The matrix of the combined similarities is as follows:

Table 3. The Combined Similarity Values

O /O'	Book	Translator	Publisher	Writer
Product	0.16	0.16	0.11	0.14
Provider	0.14	0.16	0.30	0.40
Creator	0.16	0.49	0.11	0.30

If we filter out the matrix of combined similarities (Table 3), with a threshold s = 0.30 we obtain the following alignment, (Provider, Publisher), (Provider, Writer), (Creator, Translator), (Creator, Writer)}. For more details about the approach see section 5.

5 The Proposed Approach

The architecture of our approach denoted PBW (Precision Based Weighting) is illustrated in Fig. 2. We have in input two ontologies O_1 and O_2 to be aligned. The ***Analysis Module*** performs the entities extraction from O_1 and O_2 using API Jena. Then, the ***Similarities Generation Module*** calculates for each pair of concepts (C, C') $\in O_1 \times O_2$ three similarity values using three techniques namely: the edit distance [5], the Jaro metric [5] and the similarity metric based on WordNet (Wu-Palmer algorithm) [5]. It should be noted at this level that the parameter object of the comparison is primarily the *aggregation method of similarity values* (the estimation of the weights to be assigned to matchers).

For that reason, we have set the same matchers for all three compared methods H, LCD (see the section 3 for the definition of these methods) and PBW in order to not have the results skewed by the choice of matchers. Therefore, the selection of the matchers has not been the subject of special attention. We have limited to the linguistic-based and string-based matchers. These similarity values are used by the ***Weights Estimation Module*** in order to calculate the confidence to be associated to the matchers mentioned above. The ***Similarities Combination Module*** generates then the combined similarity values using a weighting summation strategy. Finally, the ***Alignment Extraction Module*** selects the final alignment. This selection is simply performed by filtering the combined similarity values on a given threshold.

The contribution of the paper lies in the combination of similarity values. We detail below the principle of the proposed approach.

The approach proposed in this paper is an aggregation approach of similarity values calculated by several matchers. It fits into the category of automatic techniques for assigning weights to matchers which estimates their importance. We give in this section its principle.

Let O and O' be two ontologies to be aligned and let $M_1,\ldots, M_k$ k matchers which execute in parallel and calculate the similarity values between entities $e_1,\ldots, e_n$ for O and $e'_1,\ldots, e'_m$ for O' respectively. Let us note $S_1,\ldots, S_k$ the similarities matrices generated by matchers $M_1,\ldots, M_k$ respectively. The problem here is to assign to each matcher M_i a weight w_i which expresses its importance in a given alignment task.

The intuition behind this approach consists to assign to the matcher M_i the weight w_i which is equal to an estimation of the precision of M_i. Indeed, as the precision metric is a good estimation of the matcher quality; we propose to use it as an estimator of the weight that will be assigned to the matcher.

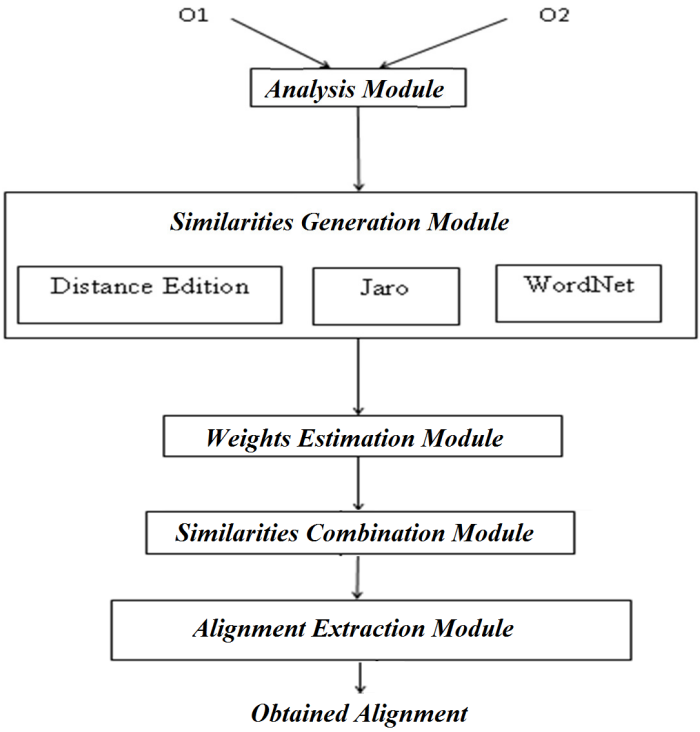


Fig. 2. The Architecture of the Application

We distinguish for the matcher M_i two subsets among the set of semantic correspondences between entities of ontologies O and O' to be aligned. On one hand we have the set P_i of the correspondences qualified positively and which belong to candidate alignment and On the other hand, we have the set N_i of those, negatively qualified and which do not belong to candidate alignment. The set P_i is defined as follows:

$$P_i = \{(e_i, e'_j) \in O \times O' / (S_i(i, j) \geq s \text{ where } s \text{ is a given threshold}\}.$$

Since to estimate the precision of a matcher, we need a reference alignment and in the absence of such alignment we propose to estimate it (the reference alignment) by the set P which denotes the set of positive correspondences identified simultaneously by all matchers. In other words: $P = \cap_{i=1,\ldots,k} P_i$. We therefore propose for the matcher M_i the following estimator for the precision: $w_i = |P_i \cap P| / |P_i|$

Where |E| denotes the number of all elements of the set E. This estimator represents the weight to be assigned to M_i.

The approach can be made operational by the following process:

– Calculate, for each matcher M_i, the set P_i defined above.

- Calculate the set P. In some alignment tasks, the case where P is empty can occur. The weights assigned to the matchers are therefore is null. To overcome this situation, we have estimated P, for each matcher M_i, as follows: $P = \{(e_i, e'_j) \in O \times O' / (S_i (i, j) \geq s$ where s is a threshold relatively high$\}$. We have retained the following formula to specify the threshold: s = the highest similarity value calculated by the matcher M_i from which a particular constant value n is subtracted. For example, if the maximum similarity value is equal to 0.80 and n=0.25 then s = 0.8*0.25=0.40.
- Assign to each matcher M_i the weight w_i.
- Calculate the matrix of combined similarity values M. the matrix M is calculated by the following formula: $M (i, j) = (\sum_k w_k * S_k (i, j)) / (\sum_k w_k)$.
- Filter M according to the threshold s.

6 The Experimental Study

In order to evaluate our approach, we have used the conference track of OAEI 2012 evaluation campaign. This track consists of a collection of 16 ontologies describing the field of the conferences organization. It is constituted of 21 tests for which reference alignments are available, from a total of 120 possible tests resulting from the pairwise combination of 16 ontologies. Each test consists of two ontologies and a reference alignment.

The tests have been carried as follows: we have implemented the three methods (H, PBW and LCD), then we have executed these methods on ontologies tests of the conference track.

As evaluation criteria we have used the standard metrics that are precision, recall and F-measure to evaluate our approach. These metrics are defined as follows:

$$\text{Precision} = P(A, R) = \frac{|R \cap A|}{|A|} \quad \text{Recall} = R(A, R) = \frac{|R \cap A|}{|R|} \quad \text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Where |R| denotes the number of the reference alignment mappings and |A| denotes the number of matches found by our approach.

We envision in this experimental analysis to compare our precision based weighting approach of the similarity values aggregation (Noted PBW-method in the graphs) with the two most efficient aggregation methods [10] [15] namely the method based on the harmony concept [8] (Noted H-method in the graphs) and the method based on the concept of local confidence [2] (Noted LCD-method in the graphs).

It should be noted at this level that our approach as well as those with which it has been compared belong to the same category of methods based on the weighting.

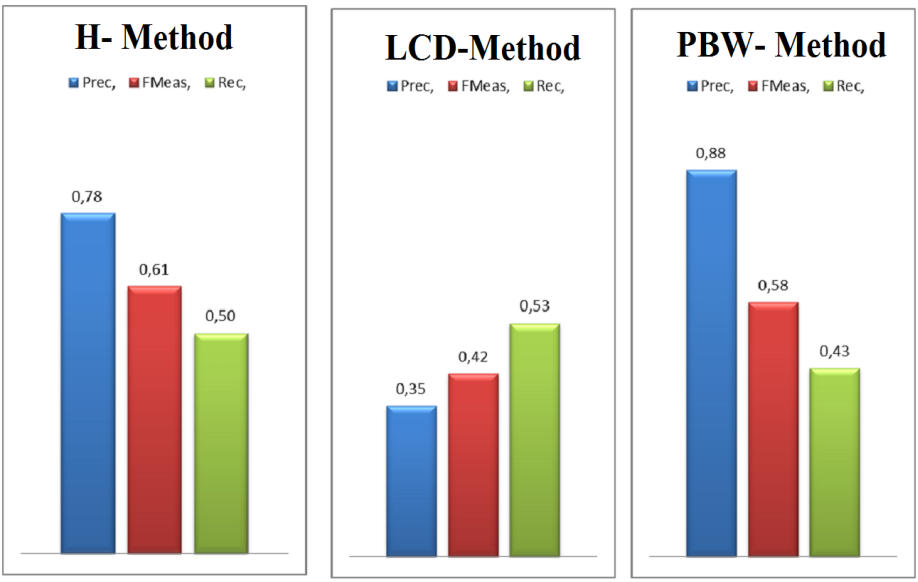


Fig. 3. The Global Results (All Tests of Conference Track) of the Three Methods

We have adopted the following methodology in order to conceive the experimental protocol. For each of the 21 tests of the conference track, we have calculated three matrices of similarities by the matchers edit distance, Jaro and WordNet, respectively. Subsequently, from these matrices we calculated the weights to assign to the matchers by the three compared methods (harmony, local confidence and our method). Then we have calculated the matrices of the combined similarities and we have selected the alignments by filtering using a given threshold s for each of the three methods. Finally, for each test we have calculated precision, recall and F-measure for each method. To conclude, we have calculated the average precision, recall and F-measure for all tests of the conference track.

The results are shown in Fig. 3 and Fig. 4.

The experimental results obtained (Fig. 3) show that globally i.e. for all tests:

1. Our approach PBW is significantly more efficient than the H and LCD methods in terms of precision.
2. Our approach PBW is more efficient than the LCD method and slightly less efficient than the H method in terms of F-measure.
3. Our approach PBW is less efficient than the H and LCD methods in terms of recall.

The analysis of the detailed results on all tests of the conference track (Fig. 4) show that our approach is more efficient than H and LCD methods in terms of precision for all tests of the conference track of OAEI 2012 evaluation campaign, considered individually.

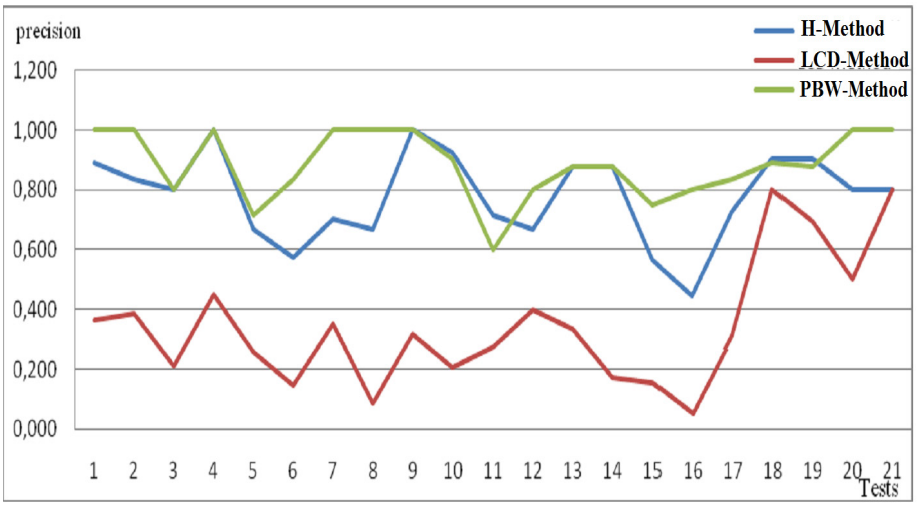


Fig. 4. The Detailed Results of the Three Methods in Terms of Precision

7 Conclusion and Perspectives

We have presented in this paper a dynamic approach to estimate automatically the weights to be assigned to different matchers in a given alignment task, in order to combine the similarity values calculated by the matchers in a context of ontology alignment.

The experimental results show the good performance of our proposed approach. Indeed, it is better in terms of precision than other methods, local and global, deemed among the most efficient ones in recent years. In addition, it shows a good F-measure relative compared to the local method.

As future perspective we envision to intensify the experiments by considering other tests and combining other similarities calculation techniques.

References

1. Bellahsene, Z., Duchateau, F.: Tuning for Schema Matching Schema Matching and Mapping. In: Bellahsene, Z., Bonifati, A., Rahm, E. (eds.) Data-Centric Systems and Applications. Springer (2011)
2. Cruz, I., Antonelli, F.P., Stroe, C.: Efficient selection of mappings and automatic quality-driven combination of matching methods. In: International Workshop on Ontology Matching (2009)
3. Do, H., Rahm, E.: COMA - A system for flexible combination of schema matching approaches. In: Proceedings of the 28th VLDB Conference, Hong Kong, China (2002)
4. Ehrig, M.: Ontology Alignment: Bridging the Semantic Gap. Springer (2007)
5. Euzénat, J., Shvaiko, P.: Ontology Matching. Springer (2013)
6. Ichise, R.: Machine learning approach for ontology mapping using multiple concept similarity measures. In: ACIS-ICIS. IEEE Computer Society (2008)

7. Li, J., Tang, J., Li, Y., Luo, Q.: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Transactions on Knowledge and Data Engineering 21 (2009)
8. Mao, M., Peng, Y., Spring, M.: A harmony based adaptive ontology mapping approach. In: Proceedings of International Conference on Semantic Web and Web Services, SWWS (2008)
9. Martinez-Gil, J., Alba, E., Aldana-Montes, J.: Optimizing Ontology Alignments by Using Genetic Algorithms. In: Gueret, C., Hitzler, P., Schlobach, S. (eds.) Nature Inspired Reasoning for the Semantic Web, CEUR Workshop Proceedings (2008)
10. Ngo, D.: Enhancing Ontology Matching by Using Machine Learning, Graph Matching and Information Retrieval Techniques. Thèse de doctorat de l'université de Grenoble (2012)
11. Rahm, E.: Towards Large-Scale Schema and Ontology Matching. Schema Matching and Mapping. In: Bellahsene, Z., Bonifati, A., Rahm, E. (eds.) Data-Centric Systems and Applications. Springer (2011)
12. Silvana, C., Ferrara, A., Montannelli, S., Varese, G.: Ontology and Instance Matching. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Multimedia Information Extraction. LNCS (LNAI), vol. 6050, pp. 167–195. Springer, Heidelberg (2011)
13. Wang, J., Ding, Z., Jiang, C.: GAOM: Genetic Algorithm based Ontology Matching. In: Proceedings of IEEE Asia-Pacific Conference on Services Computing (2006)
14. Wang, R., Wu, J., Liu, L.: Strategies Prediction and Combination of Multi-strategy Ontology Mapping. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 220–227. Springer, Heidelberg (2010)
15. (Site 1), `http://oaei.ontologymatching.org/2011/results/oaei2011.pdf` (accessed January 2015)

Exact Reasoning over Imprecise Ontologies

Mustapha Bourahla(✉)

Computer Science Department, University of M'sila,
Laboratory of Pure and Applied Mathematics (LMPA),
BP 166 Ichebilia, M'sila 28000, Algeria
mbourahla@hotmail.com

Abstract. A real world of objects (individuals) is represented by a set of assertions written with respect to defined syntax and semantics of description logic (formal language). These assertions should be consistent with the ontology axioms described as terminology of knowledge. The axioms and the assertions represent ontology about a particular domain. A real world is a possible world if all the assertions and the axioms over its set of individuals, are consistent. It is possible then to query the possible world by specific assertions (as instance checking) to determine if they are consistent with it or not. However, ontology can contain vague concepts which means the knowledge about them is imprecise and then query answering will not possible due to the open world assumption if the necessary information is incomplete (it is currently absent). A concept description can be very exact (crisp concept) or exact (fuzzy concept) if its knowledge is complete, otherwise it is inexact (vague concept) if its knowledge is incomplete. In this paper we propose a vagueness theory based on the definition of truth gaps as ontology assertions to express the vague concepts in Ontology Web Language (OWL2) (which is based on the description logic SROIQ(D)) and an extension of the Tableau algorithm for reasoning over imprecise ontologies.

Keywords: Vagueness · Ontology · OWL · Description logics · Automatic reasoning

1 Introduction

Formalisms for dealing with vagueness have started to play an important role in research related to the Web and the Semantic Web [7,13]. Ontologies are the definition of domain concepts (extensions) and the relations between them. Formal ontologies are expressed in well-defined formal languages (for example, OWL2) [3,6] that are based on expressive description logics (for example, SROIQ(D)) [1,14,4]. We say ontology is vague if it has at least a vague definition of a concept. A concept (an extension) is vague if it defines a meaning gap with which we cannot decide the membership of certain objects (vague intension).

We state the problem with the following example. Assume an ontology defining a concept called *Expensive* in a domain about cars. The meaning of the concept is vague. This vagueness is pervasive in natural language, but until now

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 355–366, 2015.
DOI: 10.1007/978-3-319-19578-0_29

is avoided in ontologies definitions. For the concept *Expensive*, we can define three sub-extensions, definitely expensive extension (there are some car prices that we regard as definitely expensive), definitely cheap extension (others we regard as definitely cheap cars) and a vagueness extension, average car prices are neither expensive nor cheap. The source of this indecision is the imprecise definition of concepts that is caused by lack of rigorous knowledge.

Related Works: Almost all concepts we are using in natural language are vague (imprecise). Therefore common sense reasoning based on natural language must be based on vague concepts and not on classical logic. The rising popularity of description logics and their use, and the need to deal with vagueness, especially in the Semantic Web, is increasingly attracting the attention of many researchers and practitioners towards description logics able to cope with vagueness. There are many works in literature for dealing with vagueness and most of them express it as a concept property as those based on fuzzy logics.

The notion of a fuzzy set proposed by Lotfi Zadeh [15] is the first very successful approach to vagueness. Fuzzy description logics (FDLs) are the logics underlying modes of reasoning which are approximate rather than exact, assertions are true to some degree [13,2,8,12]. In this case, any concept instance will have a degree of membership that is determined by a defined fuzzy function. The vagueness under fuzzy theory is treated by extended fuzzy description logics that are supported by fuzzy semantics and fuzzy reasoning. The fuzzy description logics are applied in many domains. The fuzzy knowledge base is interpreted as a collection of constraints on assertions. Thus, the inference is viewed as a process of propagation of these constraints.

Assertions in fuzzy description logics, rather being satisfied (true) or unsatisfied (false) in an interpretation, are associated with a degree of truth using semantic operators, where the membership of an individual to the union and intersection of concepts is uniquely determined by its membership to constituent concepts. This is a very nice property and allows very simple operations on fuzzy concepts. In addition to the standard problems of deciding the satisfiability of fuzzy ontologies and logical consequences of fuzzy assertions from fuzzy ontologies, two other important reasoning problems are the best truth value bound problem and the best satisfiability bound problem.

In our work, the concepts are treated as having a fixed meaning (not a balanced meaning), shared by all users of the ontology; we propose instead that the meaning of a vague concept evolves during the ontology evolution, from more vague meaning to less vague meaning until it reaches if possible, a situation where it becomes non-vague concept. This meaning instability is the base of our vagueness theory that is used for reasoning over vague ontologies. Both theories represent two different approaches to vagueness. Fuzzy theory addresses gradualness of knowledge, expressed by the fuzzy membership, whereas truth gap theory addresses granularity of knowledge, expressed by the indiscernibility relation. The result of reasoning over vague ontology using truth gap theory is the posterior description that represents a revision of the prior description on the light of the evidence provided by acquired information. This property can be

used to draw conclusions from prior knowledge and its revision if new evidence is available.

The other closest work to ours is the work in [10] which presents a framework for adjusting numerical restrictions defining vague concepts. An inconsistency problem can happen when aligning the original ontology to another source of ontological information or when ontology evolves by adding learned axioms. This adjustment is used to repair the original ontology for avoiding the inconsistency problem by modifying restrictions parameters called adaptors specified as concept annotations. The idea of this work is close to ours in the sense that we reduce the truth gaps when adding new assertions as learned knowledge to the ontology to guide the reasoning process which will play the same role as adjusting the vague concept restrictions. However, this work differs from our approach by the repair (modification) process applied on the original ontology to avoid introduced inconsistency. In our approach, we define the vague concepts as super concepts over restriction definitions. So, we don't have the problem of inconsistency to repair the ontology.

This paper is organized as follows. We begin in Section 2, by presenting Ontology Web Language (OWL2) and its correspondent description logic (SROIQ(D)). In Section 3, a vagueness theory is proposed to show how to express vague concepts and to describe the characteristics of vague ontologies. Section 4 presents the extended version of Tableau algorithm, to reason over imprecise ontologies. At the end, we conclude this paper by conclusions and perspectives.

2 Description Logics and Ontology Web Language

Ontologies are definitions of concepts and the relationships between them. They can be represented formally using formal languages. These formal description languages are based on well-defined Description Logics (DLs) [1], a family of knowledge representation formalisms. OWL2 DL is a variant of SROIQ(D) [4], which consists of an alphabet composed of three sets of names. The set $\mathcal{C}$ of atomic concepts corresponding to classes interpreted as sets of objects, the set $\mathcal{R}$ of atomic roles corresponding to relationships interpreted as binary relations on objects and the set $\mathcal{I}$ of individuals (objects). It consists also of a set of constructors used to build complex concepts and complex roles from the atomic ones. The roles (object or concrete) are called properties; if their range values are individuals (relation between individuals) then they are called object (abstract) properties. If their range values are concrete data (relation between individual and a concrete data) then they are called data (concrete) properties. The set of SROIQ(D) complex concepts can be expressed using the following grammar:

$$\begin{aligned} C ::= \top \mid \bot \mid A \mid \{a\} \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid \exists o.Self \mid \forall o.Self \mid \\ \exists o.C \mid \forall o.C \mid \exists c.P \mid \forall c.P \mid \geq n\ s.C \mid \leq n\ s.C \end{aligned} \tag{1}$$

Where $\top$ is the universal concept, $\bot$ is the empty concept, A is an atomic concept, a an individual, C and D are concepts, o an object role, c a concrete

role, s a simple role w.r.t. $\mathcal{R}$, and n a non-negative integer. P is a predicate over a concrete domain that can have the form

$$P ::= DataType\,[\sim value] \mid P \sqcap P \mid P \sqcup P, \quad \sim \in \{<, \leq, >, \geq\} \tag{2}$$

The data type can be any recognized data type as integer, real, etc. This syntax allows expressing concepts and roles with a complex structure. However, in order to represent real world domains, one needs the ability to assert properties of concepts and relationships between them. The assertion of properties is done in DLs by means of an ontology (or knowledge base). A SROIQ(D) ontology is a pair $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$, where $\mathcal{T}$ is called a terminological box and $\mathcal{A}$ is called an assertional box. The terminological box consists of a finite set of axioms on concepts and roles. There are inclusion axioms on concepts, object and concrete roles to define a hierarchy (taxonomy) on the names of concepts and roles, (we write $C \sqsubseteq D$ to denote inclusion axioms on concepts, where C and D are concepts, $C \equiv D$ as an abbreviation for $C \sqsubseteq D \wedge D \sqsubseteq C$ and $r_1 \sqsubseteq r_2$ for role inclusion, where r_1 and r_2 are object (concrete) roles, the same equivalence abbreviation can be applied on roles). The assertional box consists of a finite set of assertions on individuals. There are membership assertions for concepts ($C(a)$ means the object (individual) a is member of C), membership assertions for roles ($o(a, b)$ means the objects a and b are related by the object property o and $c(a, d)$ means the object a has the data property (concrete role) c with a value equals d).

Thus, the assertional box $\mathcal{A}$ of a knowledge base, provides a description of a world. It introduces individuals by specifying their names, the concepts to which they belong, and their relations with other individuals. The semantics of the language uses either the closed world assumption or the open world assumption. With the closed world assumption, we consider that the world is limited to what is stated. It is this assumption that is normally adopted in databases. In description logics, it is rather the assumption of the open world which prevails. This open world assumption has an impact in the way of making inferences in description logics. The inference is more complex with the assumption of the open world; it is often called to consider several alternative situations for the proof. Another important aspect of description logic is that it does not presuppose the uniqueness of names (the standard names). That is, two different names do not necessarily mean that there is case to two separate entities in the described world. To be sure that two different entities a and b are represented, should be added the assertion $a \neq b$ to the assertional box $\mathcal{A}$.

3 Vagueness Theory for Imprecise Ontologies

We define a concept C as vague if it has a deficiency of meaning. Thus, the source of vagueness is the capability of meaning (it has borderline cases). For example, the concept *Expensive* is extensionally vague and it remains intentionally vague in a world of expensive and non-expensive cars. This means that there are truth-value gaps where a vague concept is extensionally (intensionally) definitely true

(tt), definitely false (ff) and true or false (tf). Let us consider the following ontology.

$$\mathcal{O} = \left\langle \mathcal{T} = \left\{ \begin{array}{l} Dom(price) \equiv \top, Rge(price) \equiv (int[\geq 0]) \sqcap (int[\leq 100]), \\ Dom(speed) \equiv \top, Rge(speed) \equiv (int[\geq 100]) \sqcap (int[\leq 300]), \\ ExpensiveCar \equiv Car \sqcap \exists price.\,(int\,[\geq 50])\,, \\ NonExpensiveCar \equiv Car \sqcap \exists price.\,(int\,[\leq 30])\,, \\ SportsCar \equiv Car \sqcap \exists speed.\,(int\,[\geq 200])\,, \\ NonSportsCar \equiv Car \sqcap \exists price.\,(int\,[\leq 150])\,, \\ ExpensiveCar \sqsubseteq Expensive, \\ NonExpensiveCar \sqsubseteq \neg Expensive, \\ SportsCar \sqsubseteq Sports, \\ NonSportsCar \sqsubseteq \neg Sports, \\ ExpensiveSportsCar \equiv Car \sqcap Expensive \sqcap Sports \end{array} \right\}, \right\rangle \quad (3)$$

$$\mathcal{A} = \left\{ \begin{array}{l} Car(a), Car(b), Car(c), Car(d), ExpensiveSportsCar(c), \\ price(a, 25), price(b, 55), price(c, 40), price(d, 45), \\ speed(a, 220), speed(b, 250), speed(c, 160), speed(d, 180) \end{array} \right\}$$

Where *price* and *speed* are two concrete roles with the universal concept as their domains and their ranges *Rge* are defined by two integer intervals. In this knowledge base (ontology), we assume the price of a definitely expensive car ($ExpensiveCar$) is greater than or equal to fifty units and it is less than or equal to one hundred units, and a definitely no-expensive car ($NonExpensiveCar$) has a price between zero and thirty units. The concept $Expensive$ and its complement are subsuming two complex concept expressions ($ExpensiveCar$ and $NonExpensiveCar$). Each concept expression contains a sub-expression that is defined as quantified (universal or existential) restriction on a concrete role (for example, the concrete role is *price* and the restricted sub-expressions are $\exists price.\,(int\,[\geq 50])$ for the concept $ExpensiveCar$ and $\exists price.\,(int\,[\leq 30])$ for the concept $NonExpensiveCar$). By the same way, we define the vague concept $Sports$ and the concept $ExpensiveSportsCar$ as a conjunction of the concepts Car, $Expensive$ and $Sports$.

We have taken advantage of the open world assumption in description logics to define vague concepts. This ontology satisfies the assertions $Expensive(b)$ and $\neg Expensive(a)$ but the assertions $Expensive(d)$ and $(\neg Expensive)(d)$ are both not satisfied. With this knowledge base (ontology), we will assign tt to $Expensive(b)$, ff to $Expensive(a)$, and tf to $Expensive(d)$. This means, there is a deficiency of meaning (truth value gaps) between $Expensive$ and $\neg Expensive$. Consequently, the concept $Expensive$ is considered vague and the same thing for the vague concept $Sports$. The assertion $ExpensiveSportsCar(c)$ is considered as acquired information to state that c is an expensive sports car in spite of the fact that the terminology does not imply this assertion from information of the object c. We will see how this acquired (learned) information will be used to decide on other instances.

Thus, the satisfaction of a membership assertion to a vague concept depends on the concrete property value and the truth gaps. The vagueness definition of a concept will create one or more truth gaps. These are convex intervals (or

ordered sequences) of values from a concrete domain with which the satisfaction of a membership assertion to the vague concept cannot be decided. There are two borderline values for each interval (or sequence). They are the lower (l) and the upper (u) bounds of a truth gap. Thus, we associate with each vague concept C a set of truth gap assertions according to a concrete role r (or to different concrete roles) used by its description.

These truth gaps assertions can be formulated using the description logic SROIQ(D) as a result of ontology description pre-processing. This will augment the ontology $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$ by the membership and property assertions to be $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \cup \{C(tt), (\neg C)(ff),\ r(x_i, l_i),\ r(y_i, u_i)\} \rangle$ if C is checked to be a vague concept according to a concrete role r, tt and ff are considered as two additional dummy individuals. The individuals x_i, y_i are either tt or ff with the conditions $x_i \neq y_i \wedge x_{i+1} = y_i$, l_i and u_i are numerical values from the range of the concrete role r with $l_i < u_i < l_{i+1}$ for $1 \leq i < n$, where n is the number of the truth gaps. This description should verify the following vagueness consistency.

Lemma 1. *(vagueness consistency). The truth gaps set defined of any vague concept C associated with a role r (or a set of roles) should verify the condition of acceptability (vagueness consistency), this means $\forall i = 1, \cdots, n-1 : x_i \neq y_i \wedge y_i = x_{i+1} \wedge l_i < u_i < l_{i+1} < u_{i+1}$. This vagueness consistency condition can be formulated using the assertions on the dummy individuals tt and ff as*

$$\begin{aligned} &(\{C(tt), r(tt, d_1), r(tt, d_2), (\neg C)(ff), r(ff, d)\} \subseteq \mathcal{A} \Rightarrow d \notin [d_1, d_2]) \wedge \\ &(\{(\neg C)(ff), r(ff, d_1), r(ff, d_2), C(tt), r(tt, d)\} \subseteq \mathcal{A} \Rightarrow d \notin [d_1, d_2]) \end{aligned} \tag{4}$$

A non-vague (crisp) concept C will have an empty set of truth gaps according to any concrete role r.

The intuition for this vagueness theory is as follows. An ontology is considered the knowledge base of an intelligent agent; if the ontology (knowledge base) $\mathcal{O}$ contains a vague concept C with respect to a concrete role r and one of its truth gaps has the smallest interval $[l, u]$, where the assertions $r(tt, u), r(ff, l)$ are in $\mathcal{O}$. The agent cannot decide if an individual (object) a with r-property value within the interval $[l, u]$ if it belongs to C or to its complement (we say that the knowledge base is incomplete). We assume that at a moment, assertions like $C(a), r(a, d)$ are added to the ontology $\mathcal{O}$, where $l < d < u$. This new information will change the ontology agent beliefs by reducing the truth gap interval to be $[l, d]$. We call that the individual a is similar to the dummy individual tt because they belong to the same concept C. Then, the individual tt will inherit the property of a. Now, if we add the assertions $(\neg C)(b), r(b, d')$ with $u > d' > d$, this will produce a vagueness inconsistency according to this vagueness theory because the agent has already changed its beliefs so that every property assertion of an individual with respect to the concrete role r where its range is greater than d should be member of the concept C. This vagueness theory is used to adjust the truth intervals (or the truth gaps) described in the original ontology by acquired new information.

3.1 Semantics for Vagueness Theory

The formal semantics of DLs is given in terms of interpretations. A SROIQ(D) interpretation is a pair $I = (\Delta^I, (.)^I)$ where Δ^I is a non-empty set called the domain of I, and $(.)^I$ is the interpretation function which assigns for every $A \in \mathcal{C}$ a subset $(A)^I \subseteq \Delta^I$, for every $o \in \mathcal{R}$ a relation $(o)^I \subseteq \Delta^I \times \Delta^I$, called object role, for every $c \in \mathcal{R}$ a relation $(c)^I \subseteq \Delta^I \times \mathcal{D}$, called concrete role ($\mathcal{D}$ is a data type as integer and string) and for every $a \in \mathcal{I}$, an element $(a)^I \in \Delta^I$. We say the interpretation I is a model of a SROIQ(D) ontology $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$, if it satisfies all the assertions in $\mathcal{T}$ and $\mathcal{A}$. In addition, it is a model of any satisfied assertion by the ontology $\mathcal{O}$. If C_r is a vague concept with respect to the concrete role r and $x_i, y_i \in \{\mathit{tt}, \mathit{ff}\}$, then the interpretation function is extended to complex concepts and roles according to their syntactic structure.

$$
\begin{aligned}
(\top)^I &= \Delta^I \\
(\bot)^I &= \emptyset \\
(\{a\})^I &= (a)^I \\
(r)^I &= (r)^I \cup \{(x_i, l_i), (y_i, u_i) \wedge x_i \neq y_i = x_{i+1} \wedge l_i < u_i,\ i = 1, \cdots, n\} \\
(C_r)^I &= (C)^I \cup \{\mathit{tt}\} \\
(\neg C_r)^I &= (\neg C)^I \cup \{\mathit{ff}\} \\
(C_r \sqcap D_r)^I &= (C_r)^I \cap (D_r)^I \\
(C_r \sqcup D_r)^I &= (C_r)^I \cup (D_r)^I \\
(\exists o.Self)^I &= \{a \in \Delta^I | \exists (a, b) \in (o)^I \wedge a = b\} \\
(\forall o.Self)^I &= \{a \in \Delta^I | \forall (a, b) \in (o)^I \Rightarrow a = b\} \\
(\exists o.C_r)^I &= \{a \in \Delta^I | \exists (a, b) \in (o)^I \wedge b \in (C_r)^I\} \\
(\forall o.C_r)^I &= \{a \in \Delta^I | \forall (a, b) \in (o)^I \Rightarrow b \in (C_r)^I\} \\
(\exists c.P)^I &= \{a \in \Delta^I | \exists (a, d) \in (c)^I \wedge P(d)\} \\
(\forall c.P)^I &= \{a \in \Delta^I | \forall (a, d) \in (c)^I \Rightarrow P(d)\} \\
(\geq n\ s.C_r)^I &= \{a \in \Delta^I |\ |\{b | (a, b) \in (s)^I \wedge b \in (C_r)^I\}| \geq n\} \\
(\leq n\ s.C_r)^I &= \{a \in \Delta^I |\ |\{b | (a, b) \in (s)^I \wedge b \in (C_r)^I\}| \leq n\}
\end{aligned}
$$

Where n is the number of truth gaps for the vague concept C, $P(d)$ means the value d verifies the predicate P and $|S|$ is the cardinality of the set S. The predefined concepts like the universal concept $\top$, the empty concept $\bot$, the atomic concepts A and the nominative concepts $\{a_1, a_2, \cdots, a_n\}$ are defined as crisp concepts and then they will not be considered as vague concepts.

Example 1. The new ontology after generation of truth gap assertions on the original ontology $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$ described in (3), is $\mathcal{O}^{new} = \langle \mathcal{T}^{new}, \mathcal{A}^{new} \rangle$ where, $\mathcal{T}^{new} = \mathcal{T}$ and using the syntax of SROIQ(D), $\mathcal{A}^{new} = \mathcal{A} \cup \{\mathit{Expensive}(\mathit{tt})$, $(\neg \mathit{Expensive})(\mathit{ff})$, $\mathit{Sports}(\mathit{tt})$, $(\neg \mathit{Sports})(\mathit{ff})$, $\mathit{price}(\mathit{ff}, 0)$, $\mathit{price}(\mathit{ff}, 30)$, $\mathit{price}(\mathit{tt}, 50)$, $\mathit{price}(\mathit{tt}, 100)$, $\mathit{speed}(\mathit{tt}, 200)$, $\mathit{speed}(\mathit{tt}, 300)$, $\mathit{speed}(\mathit{ff}, 100)\}$, $\mathit{speed}(\mathit{ff}, 150)\}$. The new interpretations with these introduced assertions are as follows. The

interpretation for the concrete role *price* is $(price)^I = \{(a, 25), (b, 55), (c, 40), (d, 45), (\mathit{ff}, 0), (\mathit{ff}, 30), (\mathit{tt}, 50), (\mathit{tt}, 100)\}$ and for the concrete role *speed*, the interpretation is $(speed)^I = \{(a, 220), (b, 250), (c, 160), (d, 180), (\mathit{ff}, 100), (\mathit{ff}, 150), (\mathit{tt}, 200), (\mathit{tt}, 300)\}$. The interpretation of the vague concept $Expensive_{price}$ is $\{b, c, \mathit{tt}\}$.

This new ontology containing concept truth gaps is considered vague and then it is incomplete for reasoning. An ontology is complete if we can assign only the definite truth values (tt and ff) to assertions. A vague (incomplete) ontology is an ontology that has at least one vague concept and then it is possible to assign the value tf to certain assertions. In addition, a vague ontology should be acceptable (Lemma 1), which means all the truth gap sets should be acceptable. We define a partial order between ontologies that is noted by $\langle \mathfrak{O}, \leq \rangle$, where $\mathfrak{O}$ is a non-empty set of ontologies describing a domain. If $\mathcal{O}_1$ and $\mathcal{O}_2$ are two ontologies from $\mathfrak{O}$ we write $\mathcal{O}_1 \leq \mathcal{O}_2$, if $\mathcal{O}_1$ is less complete than $\mathcal{O}_2$ (we say also that $\mathcal{O}_2$ extends $\mathcal{O}_1$). The relation $\leq$ (we call it also the extension relation) is based on comparison of truth gaps and it is transitive and antisymmetric. By this partial order definition, there is a canonical normal ontology $\mathcal{O}_n$ that is the least complete ontology, which can be extended by other complete ontologies.

The set $\mathfrak{O}$ has a base ontology that corresponds to description of which all other descriptions are extensions. This base ontology is composed of the terminological assertions and eventually some membership assertions. A condition that can be imposed on domain ontology is its completeability. It states that any intermediate ontology can be extended to a complete ontology. We suppose that ontology $\mathcal{O}$ has a vague concept C, with an acceptable set of truth gaps defined by the assertions set $\{C(\mathit{tt}), (\neg C)(\mathit{ff}), r(x_1, l_1), r(y_1, u_1), r(x_2, l_2), r(y_2, u_2), \cdots, r(x_n, l_n), r(y_n, u_n)\}$, then we define the ontology extension ($\oplus$) by the assertions $\{C(a), r(a, d)\}$ as follows.

$$\langle \mathcal{T}, \mathcal{A}\,[C(\mathit{tt}), (\neg C)(\mathit{ff}), \cdots, r(x_i, l_i), r(y_i, u_i), \cdots] \rangle \oplus \langle \mathcal{T}, \{C(a), r(a, d)\} \rangle = \left\langle \mathcal{T}, \mathcal{A} \cup \{C(a), r(a, d)\} \cup \left\{ \begin{array}{l} r(x_i, d)\ if\ x_i = \mathit{tt} \wedge l_i < d < u_i \\ r(y_i, d)\ if\ y_i = \mathit{tt} \wedge l_i < d < u_i \end{array} \right\} \right\rangle$$

$$\langle \mathcal{T}, \mathcal{A}\,[C(\mathit{tt}), (\neg C)(\mathit{ff}), \cdots, r(x_i, l_i), r(y_i, u_i), \cdots] \rangle \oplus \langle \mathcal{T}, \{(\neg C)(a), r(a, d)\} \rangle = \left\langle \mathcal{T}, \mathcal{A} \cup \{(\neg C(a)), r(a, d)\} \cup \left\{ \begin{array}{l} (x_i, d)\ if\ x_i = \mathit{ff} \wedge l_i < d < u_i \\ r(y_i, d)\ if\ y_i = \mathit{ff} \wedge l_i < d < u_i \end{array} \right\} \right\rangle$$

This extension guarantees the ontology stability if the acquired informations are satisfied by the ontology description to be extended.

Lemma 2. *(stability property of $\langle \mathfrak{O}, \leq \rangle$). Let α be an assertion, we say $\langle \mathfrak{O}, \leq \rangle$ is stable if*

$$\forall \mathcal{O}_1, \mathcal{O}_2 \in \mathfrak{O}, \mathcal{O}_1 \leq \mathcal{O}_2 : \mathcal{O}_1 \models \alpha \Rightarrow \mathcal{O}_2 \models \alpha \text{ and } \mathcal{O}_2 \not\models \alpha \Rightarrow \mathcal{O}_1 \not\models \alpha$$

The complete ontology may not be available to remove completely the vagueness, thus it is necessary to work with the most extended ontology. This means, the truth-valuation is based upon the most extended ontology. Ontology can be extended to complete ontology by learned assertions as a process of ontology evolution when using an intelligent agent or inferred assertions. The learned assertions can be imported from other domain ontologies, RDF databases or simply added by the user. In the following, we propose an extension of reasoning that can take into account the proposed vagueness theory.

4 Reasoning over Imprecise Ontologies

An interpretation I is a model of an ontology $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$ denoted by $I \models \mathcal{O}$ if I satisfies all the axioms in $\mathcal{T}$ and all the assertions in $\mathcal{A}$. The reasoning is for checking concept and role instances and for query answering over a satisfiable ontology [9,11]. Ontology satisfiability is to verify whether ontology $\mathcal{O}$ admits at least one model where consistency properties should be verified. Concept instance checking is to verify whether an individual a is an instance of a concept C in every model of $\mathcal{O}$, i.e., whether $\mathcal{O} \models C(a)$. Role instance checking is to verify whether a pair (a, b) of individuals is an instance of a role r in every model of $\mathcal{O}$, i.e., whether $\mathcal{O} \models r(a, b)$.

The satisfaction properties will be extended to deal with the vagueness in ontologies. A vague (imprecise) ontology is satisfiable if it generates acceptable truth gaps for all its concepts (note that an empty set of truth gaps is acceptable). For example, if we modify the concept $ExpensiveCar$ in the vague ontology of the previous example to be $Car \sqcap \exists price.\,(int\,[\geq 55]) \sqsubseteq \neg Expensive$, this will change the set of truth gaps assertions associated with the vague concept $Expensive$ to be $\{Expensive(tt), (\neg Expensive)(ff), price(tt, 50), price(tt, 100), price(ff, 0), price(ff, 55)\}$. This set of truth gaps is not acceptable because it is a false set of assertions according to the vagueness consistency stated in (4). Nevertheless, the vague ontology is satisfiable by using the traditional reasoning techniques. However, if we add the assertions $\{Car(d), price(d, 52)\}$ to the assertional box, the vague ontology becomes inconsistent because d is now at the same time expensive car and no-expensive car, although the ontology was initially satisfiable. In the following, we will extend the reasoning Tableau algorithm to cope with the problem of vague ontologies using this proposed vagueness theory.

The principle of this reasoning algorithm is the expansion of a finite configuration $T = \{A_1, \cdots, A_n\}$ of assertions that is represented as a set of subsets, each subset is composed of assertions on individuals, using well defined rules until no rule can be applied on at least one subset (satisfaction) or contradictions (clashes) are observed within all subsets (unsatisfaction). We will have a clash in a subset A_i when a contradiction happens in it. There are three types of contradictions: $\bot(a) \in A_i$, $C(a) \in A_i \wedge (\neg C)(a) \in A_i$, or unacceptable truth gaps assertions. If no expansion rule can be applied in A_i we say that A_i is open. The terminological box should be normalized to apply the expansion rules. It is necessary to begin the inference with formulas that are independent from any

terminology. This means elimination of the definitions (equivalence axioms) and subsumptions (inclusion axioms) in the terminological box. If it contains no cycle in the definitions (which will be the case most of the time), it will happen simply by replacing all the terms in the formula by their definitions in the terminology. Obviously, if a term of formula has no definition in terminology, it remains unchanged. We repeat this process until the resulting formula contains no term which has a definition in the terminology.

For reasoning over vague ontologies using the proposed vagueness theory, we have added the following two expansion rules that should be applied after every expansion by a classical Tableau rule (the reader can be referred to [4,5,9] for the classical Tableau rules). We will get a clash (contradiction) if any new set of truth gaps assertions is not acceptable (Lemma 1 and Equation 4). The configuration length depends on ontology description and property being checked. Using the DL syntax of SROIQ(D), these two rules can be formulated as

$$V - Rule^{+}(DL) : \frac{A_i \in T \wedge \{C(a), r(a,d), C(\text{tt})\} \subseteq A_i}{(T \setminus A_i) \cup (A_i \cup \{r(\text{tt},d)\})} r(\text{tt},d) \notin A_i$$

$$V - Rule^{-}(DL) : \frac{A_i \in T \wedge \{(\neg C)(a), r(a,d), (\neg C)(\text{ff})\} \subseteq A_i}{(T \setminus A_i) \cup (A_i \cup \{r(\text{ff},d)\})} r(\text{ff},d) \notin A_i$$

These two rules will augment the assertions subset A_i by the property assertion $r(\text{tt},d)$ if A_i contains the assertion $C(a) \wedge r(a,d) \wedge C(\text{tt})$ (the rule $V - Rule^{+}(DL)$) or by the property assertion $r(\text{ff},d)$ if A_i contains the assertion $(\neg C)(a) \wedge r(a,d) \wedge (\neg C)(\text{ff})$ (the rule $V - Rule^{-}(DL)$). We explain this algorithm extension on a simple example of an instance checking using the ontology described in (3). We want to check the membership of the individual d (instance checking) to the class $ExpensiveSportsCar$ ($\mathcal{O} \models ExpensiveSportsCar(d)$). This means that we want to prove that $(\neg ExpensiveSportsCar)(d)$ is inconsistent with the ontology description. After elimination of terminological axioms and normalization as preliminary steps before applying Tableau Rules, we have:

$$T^0 = \left\{ A_0^0 = \left\{ \begin{array}{l} ((\neg Car) \sqcup (\neg Expensive) \sqcup (\neg Sports))(d), Car(a), Car(b), \\ Car(c), Car(d), price(a,25), price(b,55), price(c,40), price(d,45), \\ speed(a,220), speed(b,120), speed(c,160), speed(d,180), \\ (\neg Expensive)(a), Expensive(b), Expensive(c), Sports(a), \\ (\neg Sports)(b), Sports(c), Expensive(\text{tt}), price(\text{tt},50), price(\text{tt},100), \\ (\neg Expensive)(\text{ff}), price(\text{ff},0), price(\text{ff},30), Sports(\text{tt}), \\ speed(\text{tt},200), speed(\text{tt},300), (\neg Sports)(\text{ff}), \\ speed(\text{ff},100), speed(\text{ff},150) \end{array} \right\} \right\}$$

Using the classical expansion rule of the disjunction, we obtain the configuration:

$$T^1 = \left\{ \begin{array}{l} A_0^1 = A_0^0 \cup \{(\neg Car)(d)\}, \\ A_1^1 = A_0^0 \cup \{(\neg Expensive)(d)\}, \\ A_2^1 = A_0^0 \cup \{(\neg Sports)(d)\} \end{array} \right\}$$

We observe a clash in the subset A_0^1 (it contains $Car(d)$ and $(\neg Car)(d)$). By applying the rules $V - Rule^+(DL)$ and $V - Rule^-(DL)$, we get:

$$T^2 = \left\{ \begin{array}{l} A_0^2 = A_0^1 = \square, \\ A_1^2 = A_1^1 \cup \left\{ \begin{array}{l} price(tt, 55), price(tt, 40), price(ff, 25), price(ff, 45), \\ speed(tt, 220), speed(ff, 120), speed(tt, 160) \end{array} \right\}, \\ A_2^2 = A_2^1 \cup \left\{ \begin{array}{l} price(tt, 55), price(tt, 40), price(ff, 25), \\ speed(tt, 220), speed(ff, 120), speed(tt, 160), speed(ff, 180) \end{array} \right\} \end{array} \right\}$$

It is clear that the subset A_1^2 of assertions contains unacceptable truth gaps assertions (the following implication $\{Expensive(tt)$, $price(tt, 40)$, $price(tt, 50)$, $(\neg Expensive)(ff)$, $price(ff, 45)\} \in A_1^2 \Rightarrow 45 \notin [40, 50]$ is false). The same thing for the subset A_2^2, where the implication $\{Sports(tt)$, $speed(tt, 160)$, $speed(tt, 200)$, $(\neg Sports)(ff)$, $speed(ff, 180)\} \in A_2^2 \Rightarrow 180 \notin [160, 200]$ is also false. Thus a clash is observed in the two subsets which makes d a member of $ExpensiveSportsCar$. The principle of this approach is as follows. Without this vagueness theory, d which has the price of 45 (greater than 30 and less than 50) and the speed of 180 (greater than 150 and less than 200) cannot be decided by the classical reasoners, as $Expensive$, $Sports$, $\neg Expensive$ and $\neg Sports$ because the definitions of $Expensive$ and $Sports$ are vague. However, the ontology contains an assertion indicating that the price 40 of c is an expensive price ($Expensive(c)$) and its speed 160 makes it a sports car; this information can help the reasoner to decide that the car d of price 45 and of speed 180 is also an expensive sports car.

5 Conclusion

In this paper, we have presented a vagueness theory to deal with the problem of ontologies containing vague concepts. The vague property (characteristic) of a concept is based in general, on certain concept data properties that may generate truth gaps. With the traditional reasoning methods, it is not possible to decide the membership of an individual (object) to a vague concept (class) if its data property is in the truth gap. Ontologies could have extension (evolution), where assertions may be added, intentionally or as result of inferences. This ontology evolution can reduce the truth gaps and then logically it will be possible to infer on previously undecided assertions. This proposed vagueness theory is used to extend the current reasoning method to take into account this vagueness notion. Implementation of this approach is one of our perspectives.

References

1. Baader, F.: What's new in description logics. Informatik-Spektrum 34(5), 434–442 (2011)
2. Bobillo, F., Delgado, M., Gomez-Romero, J., Straccia, U.: Joining gödel and zadeh fuzzy logics in fuzzy description logics. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20(4), 475–508 (2012)

3. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (eds.): OWL 2 Web Ontology Language: Primer. W3C (2009)
4. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible SROIQ. In: Proc. of the 10th Int. Conf. on Principles of Knowledge Representation and Reasoning, KR 2006, pp. 57–67. AAAI Press (2006)
5. Horrocks, I., Sattler, U.: A tableau decision procedure for SHOIQ. Journal of Automated Reasoning 39(39–3), 249–276 (2007)
6. Krötzsch, M.: OWL 2 profiles: An introduction to lightweight ontology languages. In: Eiter, T., Krennwallner, T. (eds.) Reasoning Web 2012. LNCS, vol. 7487, pp. 112–183. Springer, Heidelberg (2012)
7. Lukasiewicz, T., Straccia, U.: Managing uncertainty and vagueness in description logics for the semantic web. J. Web Sem. 6(4), 291–308 (2007)
8. Lukasiewicz, T., Straccia, U.: Description logic programs under probabilistic uncertainty and fuzzy vagueness. Int. J. Approx. Reasoning 50(6), 837–853 (2009)
9. Lutz, C., Milicic, M.: A tableau algorithm for DLs with concrete domains and GCIs. Journal of Automated Reasoning 38(1–3), 227–259 (2007)
10. Pareti, P., Klein, E.: Learning vague concepts for the semantic web. In: Porc. Joint WS on Knowledge Evolution and Ontology Dynamics. In Conj. with ISWC 2011, vol. 784, CEUR workshop proceedings (2011)
11. Pérez-Urbina, H., Horrocks, I., Motik, B.: Efficient query answering for owl 2. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 489–504. Springer, Heidelberg (2009)
12. Stefan, B., Peñaloza, R.: Consistency reasoning in lattice-based fuzzy description logics. Int. J. Approx. Reason (2013)
13. Straccia, U.: Foundations of Fuzzy Logic and Semantic Web Languages. CRC Studies in Informatics Series. Chapman & Hall (2013)
14. Turhan, A.-Y.: Introductions to description logics – A guided tour. In: Rudolph, S., Gottlob, G., Horrocks, I., van Harmelen, F. (eds.) Reasoning Weg 2013. LNCS, vol. 8067, pp. 150–161. Springer, Heidelberg (2013)
15. Zadeh, L.A.: Knowledge representation in fuzzy logic. IEEE Transactions on Knowledge and Data Engineering 1(1), 89–100 (1989)

Defining Semantic Relationships to Capitalize Content of Multimedia Resources

Mohamed Kharrat(✉), Anis Jedidi, and Faiez Gargouri

MIRACL - Multimedia, InfoRmation systems and Advanced Computing Laboratory,
University of Sfax, Sfax, Tunisia
med_khr@yahoo.fr, anis.jedidi@isimsf.rnu.tn,
faiez.gargouri@isimsf.rnu.tn

Abstract. Existing systems or architectures hardly provide any way to localize sub-parts of multimedia objects (e.g. sub regions of images, persons, events…), which represents hidden semantics of resources. To simplify and automate discovering hidden connections between such resources, we describe and evaluate in this paper, an algorithm for creating semantic relationships between multimedia news resources, giving a contextual schema (represented in RDF) as a result. This latter, which could eventually be used under any retrieval system, is integrated in our main multimodal retrieval system.

We have also proposed and introduced a special measure of accuracy since evaluation relies on users' intentions. An experimental evaluation of our algorithm is presented, showing encouraging results.

Keywords: Semantic · Multimedia · Relationships

1 Introduction

Semantic web is one of the most important challenges in web realm which has been subject of many researches in recent years.

One of the main elements for processing semantic web is to go further knowledge and annotation. In fact, developing semantic retrieval systems, needs information extraction, harvesting knowledge and various methods of data.

Advances in multimedia technologies have made possible the storage of huge multimedia documents collections on computer systems. But the lack of efficiency is perceived as the main obstacle for a large-scale deployment of semantic technologies. In order to allow an efficient exploitation of these collections, designing tools for accessing data resources is required.

One of the biggest challenges is the exploitation of these collections, particularly hidden or non-exploitable relations as well as search and querying. To address this problem, we propose a mechanism to generate a new defined set of hidden semantic relationships between multimedia documents.

Within the same framework, our main system [5] proposes to retrieve multimedia documents using a multimodal approach. The main characteristic of our system is the

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 367–378, 2015.
DOI: 10.1007/978-3-319-19578-0_30

use of two languages, XQuery and SPARQL to query the description of multimedia resources. Performance of our system can be significantly increased by using a semantic relationship contextual schema (see Figure1) for semantic relationships, by applying rules through an algorithm described in this paper.

The importance of discovering such links is essentially for retrieving relevant hidden resources in results.

The algorithm which is exposed here, allows the generation and publication of linked data from metadata. Any resource which is composed of many parts could or could not have many relationships with other resources.

In addition, we introduce a special measure that allows a user to rate the correctness of each relationship and penalize irrelevant ones based on its own perception.

The aim of this paper is to present how to implement semantic relationships between data along with multimedia news resources to enhance our ability to ”understand” those latter ones. In fact, news descriptive meta-data available for users, are difficult to learn about their content and capabilities, this is why we are seeking for strengthening by establishing this feature to our main system.

The next step will be the integration of this mechanism over XQuery language, which gives the possibility to add new relationships through queries. This means, we will create new relationships based on the resources which are the results of queries. In fact, XQuery will be first used to build semantic relationships over queries based on functions, and secondly, it will be used to harness them .

The last part of the paper is structured as follows. Section 2 provides an overview of closely related work. We present proposed semantic relationships and a complementary inference reasoning to build these relationships in section 3. We evaluate our approach in section 4 and finally we conclude in section5.

2 Related Work

As far as we know, there are no other works reported addressing the task of creating semantic relationships from XML content of multimedia resources. Then, we are going to briefly present here some previous studies which are quite close to our work, consisting mainly on automatic identification of relationships from unstructured documents or the use of lexical patterns for relations discovery between concepts. These have the advantage of the simplicity of collecting training corpora automatically.

While authors use a graph in [6] to model relationships between phrases inside semantic corpus Wordnet using numbers, our method does not use a graph.

Authors in [7] Establish missing semantic relations between Wikipedia entities by discovering automatically the missing entity links in Wikipedia infoboxes, which are important for creating RDF links between DBpedia instances.

Several other approaches have been proposed having various methods and techniques. Solution proposed in [3], identify semantic links between persons, products, events and other entities from Twitter based on entities topics and their types according to time axis.

In [2], authors compute concept-concept relatedness and concept-category relatedness based on heuristics by Category links and related links in Wikipedia.

In [4], author proposes RelFinder: an interactive discovery of relationships between DBpedia objects which is controlled by users, combined with the automatic mechanisms according to topological and semantic dimensions.

Our problem shares some resemblance with works in[1], where author creates new links in precise region on images. This region represents the most relevant part in XML document of each image using hierarchical structure and adds weights for every link. The goal is to ameliorate the image retrieval in the semi-structured documents.

In contrast to our approach, all these works treat mono-media documents. They propose descriptions which allow establishing relations between annotated concepts, resources and parts of resources. These works do not take into account multimedia resources and the set of sturdy relations which are between resources or between parts of a same resource. In addition, most of them, do not consider the semantic side.

In the following section, we introduce relations and rules which allow us to extract semantics from multimedia resources.

3 Semantic Relationships

We introduce here, a contextual schema which constitutes formalism for semantic relationships representation. It expresses meaning in a form that is both logically precise and humanly readable. This schema is implemented to be used in our multimodal system and represented using RDF.

The basic assumption underlying our approach is textual descriptions of resources always hide semantics that cannot always be discovered notably between concepts. Besides, meaning of some data is sometimes either unknown, ambiguous or implicit.

However, not all semantically related concepts are interesting for end users. In this paper, we have identified a number of semantic relations.

Media fragments are really parts of a parent resource. The use of identifiers seems therefore appropriate to specify these media fragments. As for any identifier, access to the parent identifier shall be possible in order to inspect its context. Providing a way being used as agreed to localize sub-parts of multimedia objects (e.g. sub regions of images, temporal sequences of videos etc.) is fundamental.

All resources in our collection are described with NewsML annotation standard for news documents. In this standard, metadata itself comes in bewildering variety. There are specific terms to describe every type of media. We harness them to extract contextual relations to be used in semantic and contextual recognition. Most visual and audio features (motion, speech, text) will be used to describe each part. For example, in order to describe the content of video news, we apply concepts to describe scenes like meeting, speech, interview, live reporting or events/topics like sports, politics and commercials. Notably, we also apply the identities of persons that can be recovered from the visual flow (person who appears on the screen), from audio or from textual information.

Our goal is to make a semantic search based on both content and structure at the same time. We do not propose to use existing links between resources, but we create our own links.

Our algorithm takes as input a resource and generates a new relationship if links exit with some other resources.

```
<?xml version="1.0"?>
<resources><resource  id="IMAGE01" type="image">
<link name= "AI">
<resource  id="IMAGE02" type="image"></resource>
<resource  id="VIDEO01" type="video"></resource>
</link>
</resource>
<resource id="VIDEO07" type="video"> <link name="SH"><resource
id="IMAGE03" type="image"></resource>
…
</link>
```

Fig. 1. Sample of Contextual Schema

3.1 Relationships Mechanism

The task of relationships building is a crossing problem between textual relations and semantic relations.

For instance, the textual expression “P talks in R” indicates a semantic relation Talk between entity “P” which is represented by a resource and another resource “R”.

This semantic relationship can be expressed textually in several ways: for example “P, said something about X” or “a quotation of P in R”.

There are several components to make a coherent relationship, including specific textual expressions as well as constraints on the entities involved in the relation. For instance, in the Talk relation, “P” must be a Person and “R” a Resource. The details of every relationship are given below.

T: Talk
This type of relationships describes links between resource R which contains {person, organization, team…} talking. This relation must be between an image and another type of document.

TA: Talk About
This type of relationships describes links between resource R which represents {document, report, documentary…} and another resource R'.

S: Speak
This type of relationships describes links between resource R which contains only a person and another resource R'.

SA: Speak About
This type of relationships describes links between resource R which contains only a person speaking, and another resource R'.

SH: Show
This type of relationships describes links between a resource R {documentary, event, interview…} which shows {person, organization, team, place…} and another resource R'.

AI: Appear In

This type of relationships describes links between a resource R which represents {person, organization, team, place…} and appears in another resource R' which represents {event, scene, sequence…}.

In the following, we briefly explain the mechanism via rules that must be used to create these relationships.

Algorithm

```
Input: Xml resource r
Output: relation between two or more resources in contex-
tual schema CS
For all r   {R} do
{Extract  metadata from r and r'
If any verified module
{ If   in CS
Add new relation to CS
End If}
Else
Execute module inference
End If
End For
Return  r ↔ r'
```

Rule 1:

$$\text{Talk}(\exists R \supset \{\text{image}\} \wedge R' \supset \{\text{video,audio,text}\} \wedge \exists a \langle objet \rangle, \langle person \rangle, or \langle ganization \rangle \supset R \supset R' \wedge \exists \{\langle interview \rangle, \langle report \rangle\} \supset R \quad (1)$$

To add the new relationship Talk, the resource origin R must be an image and the destination R' could be any type (video, audio, text). Secondly metadata like ‹object›, ‹person› or ‹organization› must exist in the two resources. In addition, ‹interview› or ‹report› must be present in the destination resource.

Rule 2:

$$\text{Speak}(\exists R \supset \{\text{image}\} \wedge R' \supset \{\text{video,audio}\} \wedge \exists \{\langle \text{person} \rangle\} \supset R \supset R' \wedge \exists \{\langle \text{interview} \rangle \vee \langle \text{speech} \rangle\} \supset R' \quad (2)$$

To add the new relationship Speak, the resource origin R must be an image and the destination R' could be (video or audio). Secondly, the metadata only ‹person› must exist in the two resources. In addition, ‹interview› or ‹speech› must be present in the destination resource.

Rule 3:

$$\text{TalkAbout}(\exists R, R' \supset \{\text{image,video,audio,text}\} \wedge R \equiv R' \wedge \text{type}(R) \neq \text{type}(R') \quad (3)$$

To add the new relationship TalkAbout, the resource origin R and the destination R' could be of any type of media (image, video, audio, text). Secondly, we fix a similarity threshold between metadata of both resources using TFIDF measure between XML's tags of these resources. The type of related resources must be different, (e.g. we could not relate two images or two videos).

Rule 4:

$$\text{SpeakAbout}(\exists R \supset \{\text{video,audio}\} \wedge R' \supset \{\text{image,video,audio,text}\} \wedge \exists \{\langle\text{person}\rangle\} \supset R \wedge R \equiv R' \quad (4)$$

To add the new relationship SpeakAbout, the resource origin R must be a video or an audio resource and the destination R' could be of any type (image, video, audio, text). Secondly, metadata ‹person› must exist in the original resource. Finally, we fix a similarity threshold between metadata of both resources using TFIDF measure between XML tags of resources.

Rule 5:

$$\text{Show}(\exists R \supset \{\text{video,image}\} \wedge R' \supset \{\text{image,video,audio,text}\} \wedge \exists \{\langle\text{documentary}\rangle, \langle\text{event}\rangle, \langle\text{interview}\rangle\} \supset R \wedge R \equiv R' \quad (5)$$

To add the new relationship Show, the resource origin R could be only a video or an image and the destination R' could be any type (image, video, audio, and text). Secondly metadata {documentary, event, interview…} must exist in the original resource. Finally, we fix a similarity threshold between metadata of both resources using TFIDF measure between XML tags of resources.

Rule 6:

$$\text{AppearIn}(\exists R \supset \{\text{image}\} \wedge R' \supset \{\text{video,image}\} \wedge \exists \{\langle\text{objet}\rangle, \langle\text{person}\rangle, \langle\text{organization}\rangle\} \supset R \supset R' \quad (6)$$

To add the new relationship AppearIn, the resource origin R must be only an image and the destination R' could be (image or video). Secondly metadata like ‹object›, ‹person› or ‹organization› must exist in the two resources.

3.2 Inference reasoning

Since XML does not support or suggest reasoning mechanisms, we have to rely on an underlying logical formalism.

We define here some inductive rules to deduce new relationships from existing relationships.

Case 1: $$\exists\ \text{link}(R1,R2) \wedge \text{proximity}(R3,R2) \Rightarrow \text{link}(R1,R3) \quad (7)$$

R1→R2 are two related resources, so if the new resource R3 has proximity with R2 then R1→R3

Case 2: $$\exists(\text{link}(R1,\{R\}) \equiv \text{link}(R2,\{R\})\ ) \Rightarrow R1 = R2 \quad (8)$$

Case 3: $$\exists(\ \text{link}(R1,R2) \wedge \text{link}(R1,R2) \wedge \text{link}(R1,R3)) \Rightarrow R2 \equiv R3 \equiv R4 \quad (9)$$

If there is semantic relationship between resource R1 and other resources as follow:

R1→R2 ; R1→R3 ; R1→R4

Then, there will be similarities between R2, R3 and R4.

We define link (R,R') as an existing semantic relation between R and R' and Proximity (R,R') as the similarity between R and R' calculated by the measure below.

3.3 Similarity Measure

We use this computation whenever a similarity measure is needed. It is composed of three steps.

First step:

Pre-processing: this module is concerned with pre-processing operations preparing the input resource to be linked. It checks if a resource R is typed.

Second step:

- Comparing ‹keyword› of R and R' for equality or similarity
- Comparing ‹title› of R and R' for equality or similarity

Third step:

Similarity is defined by some functions:

The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

In addition, we use term frequency. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j. Thus we have the term frequency, defined as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{11}$$

where $n_{i,j}$ is the number of occurrences of the considered term (t_i) in document d_j, and the denominator is the sum of number of occurrences of all terms in document d_j, that is, the size of the document $| d_j |$.

A threshold parameter is used here and changes during evaluation.

In the main system, queries attempt to find semantic contents such as specific people, objects and events in a broadcast news collection. We define the following classes: Named person, Named object and General object.

Our retrieval system needs to go through the following steps to find relevant multimedia resources for content-based queries without any user feedback or manual query expansion.

4 Evaluation

In this section we present the results of the experimental evaluation that we have conducted on the semantic relationships extraction using real datasets. The objective is to evaluate the efficiency of the schema transformation. We have used several resources sets attempting to cover many domain variations, features and special cases.
The algorithm has been implemented using PHP language on the top of the open source, native XML database.

4.1 Accuracy Evaluation

We present the performance of the schema transformation generation processes of previous sections, then we examine the efficiency of the said processes.

The experiments were conducted on a 2.53GHz Intel Core I5 machine with 4GB of RAM, running MS 8. All results are averaged over three runs.

The basic characteristics (e.g., number of elements, attributes, etc.) of the XML resources are not shown in the evaluation. In this experiment, we have chosen a context about international politics and United States politics mainly but not exclusively. We have also used various contexts, as sports, terrorism, or Internet privacy.

Annotation has an important role here, there is a consequence and a difference between, for example, the keyword “Angela Merkel” appearing or not in “keyword” tag, and/or in “subject” tag too. Besides, taking for example, the keyword Hillary Clinton which could exist in other forms like Hillary Rodham Clinton or even Hillary Diane Rodham Clinton, has a real impact on results. This depends on news agencies, but actually, we did not deal with this point.

Table 1 shows results by number of used resources. Even a small number of resources is used, we believe that the use of a huge database could not have an impact on the results or on the performances. The result of the variation of used thresholds is presented in table 2. We can observe that by increasing it, redundancy decreases but the total number found decreases too.

Finally, as expected, the XML Schema file size slightly affects the time consumed by this transformation, e.g. a single iteration for a result set size of 20 resources takes about 3 seconds.

Table 1. Results values by modifying the number of resources

Resources	Total	Erroneous	Valid	Not detected
5	5	2	3	2
20	8	2	6	3
30	11	3	8	5

Table 2. Results values using variable thresholds

	Total	Not found	Redundancy
Similarity threshold >1	8	3	1
Similarity threshold >1.5	7	3	0

Table 3. Sample of the results values by relationship types

Relationship	Detected	Erroneous	Not detected
Show	1		1
Appear In	2	1	1
Talk	0		
Talk About	1		
Speak	1	1	
Speak About	3		1

In table 3, we present the results by relationship types. We assume and believe that the contents are the primary determinants of these results.

In fact, there is a lot of data which are made by humans, so the same content could be written in many ways hence influencing the interpretation even if made by humans. Consequently, this can result in weak structures. The fact of omitting Named Entity or describing differently a situation could change retrieval results.

Also the fact of using resources with the same context is very important, because otherwise, we could have zero relationships. Finally, the rank of metadata in XML resources is also computed and has an effect on results too.

e.g.:

<genre qcode="genre:WarConflict"></genre>
<genre qcode="genre:Politics"></genre>

These two tags do not have the same impact if the creator of the metadata considers the importance of the rank of those genres or chooses to put the inverse.

We note that we have more relationships between the same type of resources than between different types even those between videos and texts. We have an interesting number of relations. We can note that the factor “type” is important.

Notice that the poor descriptions of images has impacts on the results because usually images are not well annotated, seeing their nature. In addition, images have no <title> tag. Besides, the image sometimes describes a general context and does not specify persons or known entities, e.g. a picture containing scenes of injury or dead bodies.

Furthermore, we would measure the Recall as the fraction of the relations that are relevant and successfully retrieved, and the Precision as the fraction of retrieved documents that are relevant to the result obtained according to our perception.

$$Precision = \frac{|Correct\ detected\ relations|}{|Total\ retrieved\ relations\ |} \quad (12)$$

Precision: a fraction of documents that are relevant among the entire retrieved document. Practically it gives accuracy of the results.

$$Recall = \frac{|Correct\ detected\ relations|}{|Total\ correct\ relations\ |} \tag{13}$$

Recall: a fraction of the documents that is retrieved and relevant among all relevant documents. Practically it gives coverage of the results.

Precision is more important than Recall in our case, because the irrelevant/wrong relationships has negative repercussions on the results in our retrieval system. Figure 2 shows that the overall precision of our system is 0.69 indeed.

Consider also that normally, the number of used resources do not affects results. Performance will be quite close to this limit even if we increase this number. In some cases, for example, in the absence of connexion between resources, we could get 0 relationships which does not imply irrelevance of the algorithm.

In the next section, we introduce a new measure more meaningful than Recall and Precision.

4.2 Accuracy Metrics

It is well known, that the policy of the user providing relevance feedback can have a strong impact on the evaluation results. Since the user's views differ, judging the correctness of the retrieved relationships is a challenging task besides the distinction between relevant combination of relationships which is related to different interpretations.

In fact, the correctness of a detected relation is not a bivalent value as it is based on the user's perception. A relation could be irrelevant or missed. In essence, for an evaluation, a missed relation is better than an irrelevant one because this latter could have repercussions on the research results.

We employ and invite three testers to evaluate how closely the results satisfied their intentions.

For every user, we compare its interpretation with the original one. To assess the correctness of the algorithm, the results were manually examined by domain experts, and for this reason, we introduced a special measure that is Alg_i.

This measure allows a user to rate the correctness of each relationship and penalize the irrelevant ones, depending on its own perception.

Let us consider the following:

i, j : indexes of documents and k: number of relationships where:

$$\begin{matrix} i,j \\ i \neq j \end{matrix} \in]0,1[^2 \qquad k \in \{1 \rightarrow 6\}$$

M : number of users

I : {indexes of all documents}

S: relations between I & j and

$$Alg_i\{L_{ij}^k\}$$

For every Alg_i we associate a subset E_{ij} (indexes of relations between fixed i and j)

$$Alg_{X1}: \{L_{ij}^{s}\}$$

Alg(x1) =C1=C1*+C1**

To compute C1, we fixe X* and X** which are penalties

To penalize (-) ↔X*

To penalize (+)↔X**

Where: X** « X*

```
01.    In case of irrelevant relationship
02.    Initializing  C1*←0
03.    For i,j ϵ I  /i≠j
04.    For k ϵ Eij
05.    For sϵ Xij and s≠k :C1*←C1*+X*
06.    Return C1*
07.    In case of missed relationship
08.    Initializing C1**←0
09.    For i,j ϵ I  /i≠j
10.    For s ϵ Xij
11.    For k ϵ Eij and s≠k :C1**←C1**+X**
12.    Return C1**
```

0<=C1<=6X*

X*ϵ]0,1[

$0<=G=\frac{C1+C2+\cdots+CM}{M}<=6X^*$

0 <= G <=1

We simply have to set X as a scalar to get results. According to our metric, the more the number tends towards zero, the more this number is relevant. In our case, X is set to 1/3. We perceive through figure 3, that all obtained results are close and good. This has an important impact on ambitious efforts to detect relationships with more efficiency.

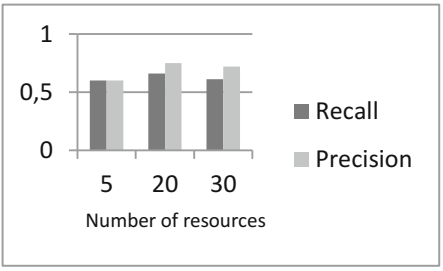


Fig. 2. Accuracy Recall & Precision

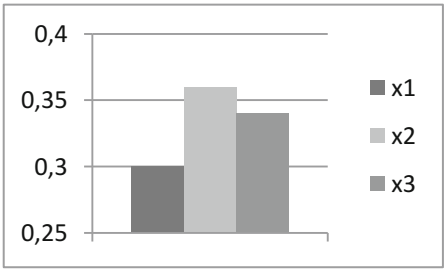


Fig. 3. Accuracy of Alg_i

5 Conclusion

Metadata provides rich semantic relationships that can be used for retrieval purposes. In order to capitalize hidden connexions and relationships between resources, we have presented in this paper a proposition for interlinking multimedia resources semantically through defined rules, and then, results are supplied as a contextual schema of RDF triples. The goal of this schema is primarily refining querying, and adding more semantics to our retrieval system.

The experimental results show that our approach can accurately find hidden relations between resources, and thus our main retrieval system will perform in a better way.

Actually, among obtained relations, there are some wrongly detected relations and some correct ones which are not detected. The next step is to continue exploring ways to improve the Precision of the construction of the relationships with poorer performance. In particular, the use of inference that may express ambiguous relationships, depending on the context, needs to be further enhanced. We also plan to try this algorithm with yet newer relationships based on resulting resources, we could build new relationships that will be used in second time.

We will continue investigating on the best combination of annotation and recommendation for using keywords to get better result.

References

1. Aouadi, H., Torjmen, M.: Exploitation des liens pour la recherche d'images dans des documents XML. In: Conférence Francophone en Recherche d'Information et Applications – CORIA (2010)
2. Bu, F., Hao, Y., Zhu, X.: Semantic Relationship Discovery with Wikipedia Structure. In: 22nd International Joint Conference on Artificial Intelligence IJCAI 2011, Barcelona, pp. 1770–1777 (2011)
3. Celik, I., Abel, F., Houben, G.-J.: Learning Semantic Relationships between Entities in Twitter. In: Auer, S., Díaz, O., Papadopoulos, G.A. (eds.) ICWE 2011. LNCS, vol. 6757, pp. 167–181. Springer, Heidelberg (2011)
4. Heim, P., Lohmann, S., Stegemann, T.: Interactive Relationship Discovery via the Semantic Web. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 303–317. Springer, Heidelberg (2010)
5. Kharrat, M., Jedidi, A., Gargouri, F.: A system proposal for multimodal retrieval of multimedia documents. In: 9th IEEE International Symposium on Parallel and Distributed Processing with Applications. ISPA, Busan-Korea (2011)
6. Stanchev, L.: Building Semantic Corpus from WordNet. In: IEEE International Conference on Bioinformatics and Biomedicine Workshops, Philadelphia (2012)
7. Xu, M., Wang, Z., Bie, R., Li, J., Zheng, C., Ke, W., Zhou, M.: Discovering Missing Semantic Relations between Entities. In: Alani, H., et al. (eds.) ISWC 2013, Part I. LNCS, vol. 8218, pp. 673–686. Springer, Heidelberg (2013)

Security and Network Technologies: Security

A Multi-agents Intrusion Detection System Using Ontology and Clustering Techniques

Imen Brahmi[1(✉)], Hanen Brahmi[1], and Sadok Ben Yahia[2]

[1] Faculty of Sciences of Tunis, Computer Science Department, Campus University, 1060 Tunis, Tunisia
imen.brahmi@gmail.com
[2] Institut Mines-TELECOM, TELECOM SudParis, UMR CNRS Samovar, 91011 Evry Cedex, France
sadok.benyahia@fst.rnu.tn

Abstract. Nowadays, the increase in technology has brought more sophisticated intrusions. Consequently, Intrusion Detection Systems (IDS) are quickly becoming a popular requirement in building a network security infrastructure. Most existing IDS are generally centralized and suffer from a number of drawbacks, *e.g.*, high rates of false positives, low efficiency, etc, especially when they face distributed attacks. This paper introduces a novel hybrid multi-agents IDS based on the intelligent combination of a clustering technique and an ontology model, called OCMAS-IDS. The latter integrates the desirable features provided by the multi-agents methodology with the benefits of semantic relations as well as the high accuracy of the data mining technique. Carried out experiments showed the efficiency of our distributed IDS, that sharply outperforms other systems over real traffic and a set of simulated attacks.

Keywords: Intrusion detection system · Multi-agents · Clustering · Ontology

1 Introduction

As far the cost of information processing and Internet accessibility is dropping, more and more organizations are becoming vulnerable to a wide variety of cyber threats. Therefore, network security is becoming a major challenge. Consequently, software tools, that can automatically detect a variety of intrusions, are of a compelling need. An *Intrusion Detection Systems* (IDS) has been of use to detect and defend intrusions more proactively in short period.

Even that IDSs have become a standard component in security infrastructures, they still have a number of signicant drawbacks [14]. Indeed, they suffer from problems of reliability, relevance, disparity and/or incompleteness in the presentation and manipulation of knowledge as well as the complexity of attacks. This fact hampers the detection ability of IDS, since it causes the generation excessive of false alarms and decreases the detection of real intrusions. In addition, most of the IDSs use centralized architectures. Unfortunately this strategy has

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 381–393, 2015.
DOI: 10.1007/978-3-319-19578-0_31

several drawbacks [4]. Indeed, the central processing node can lead to a single point of failure. Clearly, whenever the central processing node is attacked, then the whole IDS has been damaged. Besides, the transfer of all the information at a central processing unit implies a great need on network resources and leads to much network load on the system. Consequently, the centralized IDS suffers from scalability problems [4]. Moreover, the communication and cooperation between a centralized IDS components are badly missing. To palliate these problems, the integration of a multi-agents technology within the IDS seemed to be an appropriate solution. In fact, the use of multi-agents system for intrusion detection offers a new alternative to the IDS with several advantages listed in literature, *e.g.*, independently and continuous running, minimal overhead, scalability, *etc.*, [4]. Therefore, multi-agent technology makes the resilience of the system strong and thus ensures its safety [6]

Alongside, the concept of ontology has emerged as a powerful method for domain knowledge representation and sharing. It can improve the intrusion detection features giving the ability to share a common conceptual understanding threats and design the signature rules [1,7,10,13,20]. In fact, the use of the ontologies and OWL (*Ontology Web Language*) within the intrusion detection context has different advantages: (*i*) Grasping the semantic knowledge about the intrusion detection subject; (*ii*) Expressing the IDS much more by building better rules of signatures using the SWRL (*Semantic Web Rule Language*) [9]; and (*iii*) Making intelligent reasoning [6,10]. In this respect, it is possible to design a multi-agents architecture based on a knowledge basis represented as an ontology. The use of such architecture reveals conducive to the development of IDSs [6].

In this paper, we investigate another way of tackling the aforementioned problems. Thus, we introduce a new distributed IDS, called OCMAS-IDS (*Ontology and Clustering based Multi-AgentS Intrusion Detection System*). OCMAS-IDS is based on the integration of the multi-agents technology, the ontology and the clustering technique. In this respect, our proposed system uses a set of agents that can be applied to a number of tasks, namely: data capturing, detecting the known and unknown attack categories and ultimately alerting the administrator. Through extensive carried out experiments on a real-life network traffic and a set of simulated attacks, we show the effectiveness of our proposal in terms of (*i*) the scalability and (*ii*) the detection ability of our system.

The remaining of the paper is organized as follows. Section 2 sheds light on the related work. We introduce our new distributed intrusion detection system based on the multi-agents technology in Section 3. We then relate the encouraging results of the carried out experiments in Section 4. Finally, Section 5 concludes and points out avenues of future work.

2 Scrutiny of the Related Work

Recently, few approaches, within the intrusion detection field, are dedicated to the integration of multi-agents technology and ontology model. Approaches fitting in the distributed IDS trend using ontological structure attempt to enhance the IDS accuracy and performing intelligent reasoning.

Worth of mention that the first research of applying ontology within intrusion detection context was done by Undercoffer et *al.* [20] in 2003. In this respect, the authors developed an ontology focused on the target (*centric*) and supply it within the format of the logical description language DARPA *DARPA Agent Markup Language + Ontology Inference Layer* (DAML + OIL). This ontology allows modeling the domain of computer attacks and facilitates the process of reasoning to detect and overcomes the malicious intrusions.

Mandujano [13] proposed a detection tool composed of a multi-agents architecture and an ontology focused on attacker, called FROID (*First Resource for Outbound Intrusion Detection*). FROID attempts to protect a set of nodes in a network using the ontology OID (*Outbound Intrusion Detection*). The proposed system is characterized by its intention to detect known attacks based on signatures. Thus, the main drawback of FORID system is that in case of an emerging attack, it will ignore it since this new attack has not yet been listed in the base of signatures.

In addition, Abdoli and Kahani [1] proposed a system, called ODIDS. The system includes two types of agents: IDSAgent and MasterAgent. Based on the techniques of the semantic web, they have built an ontology for extracting semantic relationships between intrusions. The main moan that can be addressed to the ODIDS system stands in the fact that the MasterAgent is a central point of failure. Hence, if an intruder can prevent it from working (*e.g.*, blocking or slowing the host where it is running), the entire system will be damaged. Another criticism of the ODIDS system is time wasting, since the system needs more time to make a connection between the MasterAgent and the IDSAgents on the network and to send and receive messages between them.

Azevedoln et *al.* [3] proposed an autonomic model, called AutoCore, which includes a set of intelligent agents as well as a domain ontology CoreSec, in order to perform intrusion detection independently. The system makes use of CoreSec as an ontology knowledge base with high-level concepts for information [3]. The agents are then responsible for enabling the analysis of network traffic and the detection of malicious activities. However, the approach does not consider the secure state which is important to judge false positive alerts and successful possibility of attacks [12].

In [7], Djotio et *al.* proposed a MONI system based on an ontology model, called NIM-COM. The MONI system includes a multi-agents IDS to achieve a distribution of the detection activities. In addition, MONI is endowed with a *Case Based Reasoning* (CBR) mechanism to learn new attacks. Even though CBR is considered as a powerful reasoning paradigm and easy to set up, it suffers from re-engineering problems [14]. This lack of flexibility of the knowledge representation is with no doubt an inherent CBR limitation.

With the same preoccupation, Isaza et *al.* [10] developed a multi-agents architecture for the detection and prevention of intrusions, called OntoIDPSMA. The representation of known attacks has been designed using a semantic model based on ontology specifying signatures and reaction rules. The authors integrated an Artificial Neural Network (ANN) technique and the clustering algorithm

K-MEANS for the identification of new attacks. However, the most significant disadvantage of ANN relies on the fact that its ability to identify an intrusion is completely dependent on the accurate training of the system, data and the methods that are used. Moreover, the configuration of an ANN is delicate and can significantly affect the results [18]. In addition, the performance of K-MEANS and its effectiveness as a method for detecting new attacks depends on the random selection of the number of initial groups. Therefore, a "bad choice" of this number will decrease the detection of actual intrusions and increase the generation of false alarms [4].

Due to its usability and importance, detecting the distributed intrusions still be a thriving and a compelling issue. In this respect, the main thrust of this paper is to propose a hybrid distributed IDS, called OCMAS-IDS, which integrates : (*i*) a multi-agents technology; (*ii*) an ontology; and (*iii*) an unsupervised clustering technique. The main idea behind our approach is to address limitations of centralized IDSs by taking advantage of the multi-agents paradigm as well as the ontological representation.

3 The OCMAS-IDS System

Agents and multi-agents systems are one of the paradigms that best fit the intrusion detection in distributed networks [4]. In fact, the multi-agents technology distributes the resources and tasks and hence each agent has its own independent functionality, so it makes the system perform work faster [6].

The distributed structure of OCMAS-IDS is composed of different cooperative, communicant and collaborative agents for collecting and analyzing massive amounts of network traffic, called respectively: SNIFFERAGENT, MISUSEAGENT, ANOMALYAGENT and REPORTERAGENT. Figure 1 sketches at a glance the overall architecture of OCMAS-IDS.

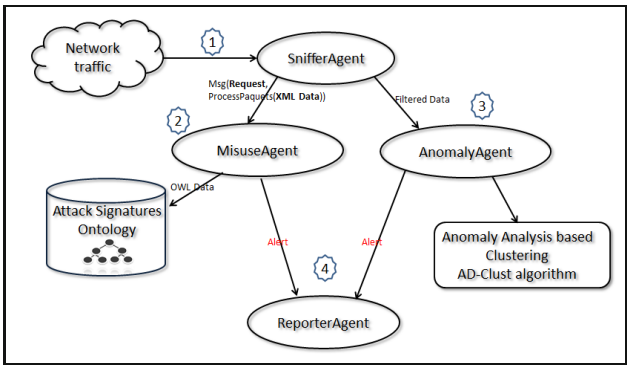


Fig. 1. The architecture of OCMAS-IDS at a glance

Worth of mention that the combination of the detection known attacks as well as the unknown ones can lead to improve the performance of the IDS and en-

hances its detection ability [11]. Consequently, OCMAS-IDS efficiently merges the detection of both types of attacks. It incorporates a MisuseAgent specialized on known attacks detection, as well as an AnomalyAgent competent on unknown attacks detection. The processing steps of OCMAS-IDS can be summarized as follows:

1. The SnifferAgent captures packets from the network. Indeed, a distributed IDS must undertake to analyze a huge volumes of events collected from different sources around the network. Consequently, the SnifferAgent permits to filter the packets already captured. Besides, it converts them to XML, using the XStream library[1]. Finally, the pre-processed packets will be sent to others agents to be analysed;
2. The MisuseAgent receives the packets converted to XML from the SnifferAgent. It transforms these packets to OWL format in order to be compatible with the SWRL rules stored in the ontology. Now, it is ready to analyze the OWL packets to detect those that correspond to known attacks. Indeed, the MisuseAgent searches for attack signatures[2] in these packets, by consulting the ontology ASO (*Attack Signatures Ontology*). Consequently, if there is a similarity between the OWL packets and the SWRL rules that define the attack's signatures, then the agent raises an alert to the ReporterAgent;
3. The filtered network packets are fed into an AnomalyAgent, which uses the clustering algorithm $\mathcal{AD}$-$\mathcal{C}$lust to detect the unknown attacks. Likewise, the agent sends an alert to the ReporterAgent, if an attack is identified;
4. Finally, the ReporterAgent generates reports and logs.

OCMAS-IDS detects the known attacks through the intelligent agent MisuseAgent, which uses an ontology to enrich data intrusions and attack signatures by semantic relationships. In what follows, we present the proposed ontology used within our system OCMAS-IDS.

3.1 The Attack Signatures Ontology (ASO)

Since last few decades, Raskin et *al.* [16] opened a new field, that focuses on using *Ontology* within information security and its advantages. In fact, ontologies present an extremely promising new paradigm in computer security domain. They can be used as basic components to perform automatic and continuous analysis based on *high-level* policy defined to detect threats and attacks [10]. Moreover, they enable the IDS with improved capacity to reason over and analyze instances of data representing an intrusion [7,20]. Furthermore, the interoperability property of the ontologies is essential to adapt to the problems of the systems distribution, since the cooperation between various information systems is supported [3,7].

[1] Available at: http://xstream.codehaus.org/ .

[2] An attack signature is a known attack method that exploits the system vulnerabilities and causes security problem [4].

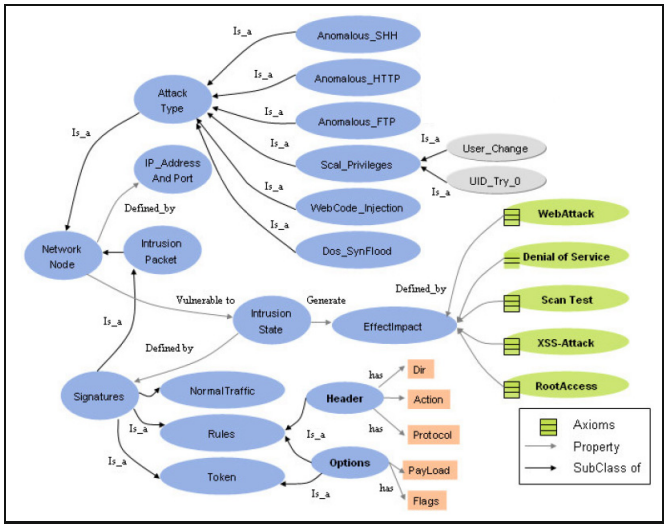


Fig. 2. The Attack Signatures based Ontology ASO

Within the OCMAS-IDS system, an ontology, called ASO (*Attack Signatures based Ontology*), is implemented, in order to optimize the knowledge representation and to incorporate more intelligence in the information analysis. Moreover, OCMAS-IDS integrates into its internal structure the interoperability between agents since they use the same model of ontology. The ASO ontology is characterized by network components, intrusion elements, classification defining traffic signatures and rules classes and instances. Figure 2 depicts a fragment of the ontology ASO, which implements the intrusion detection knowledge. The ASO ontology allows the representation of the signatures basis for known attacks, used with the agent MisuseAgent. The power and usefulness of ontology, applied to the signature basis issue, provide a simple representation of the attacks expressed by the semantic relationships between intrusion data. We can also infer additional knowledge about intrusion due to the ability of the ontology to infer new behavior by reasoning about data. Therefore, this fact improves the process of decision support for an IDS [1,6,20].

The signature basis incorporates rules provided by the ASO ontology, that allows a semantic mean for reasoning and inferences. In fact, the rules are extracted using the SWRL language (*Semantic Web Rule Language*). The latter extend the ontology and enriches its semantics by the deductive reasoning capabilities [9]. It allows to handle instances with variables (?x, ?y, ?z). Thus, the SWRL rules are developed according to the scheme: *Antecedent* $\rightarrow$ *Consequent*, where both antecedent and consequent are conjunctions of atoms written $a_1 \wedge ... \wedge a_n$. Variables are indicated using the standard convention of prefixing them with a question mark (*e.i.*, "?x"). The following example shows a rule represented with SWRL.

Example 1. NetworkHost(?z) ∧ IntrusionState(?p) ∧ GeneratedBY(?p,?z) ∧ SQLInjection(?p) ∧ Directd_To(?p,?z) → SystemSQLInjectionState(?p,?z)

Using this syntax, a rule asserting that the composition of the network host(z) and an intrusion state(p) properties implies the attack "*SQL Injection*" property.

When constructing our ontology, we designed and implemented multiple rules to define various attacks and signatures. The defined rules allow properties inferences and reasoning process. The attack properties, *e.g.*, *WebAttack*, *SQLInjection*, *DoS*, *dDoS*, and so on, are defined as ontology' attributes identifying the type of an intrusion.

Even thought, the known attacks are detected, it remains nevertheless the problem of the new attacks detection. In this respect, additionally to the MisuseAgent, based on the ontology, OCMAS-IDS uses an AnomalyAgent based on the clustering analysis. The algorithm is described in the following subsection.

3.2 The Clustering Algorithm AD-Clust

Needless to remind that the application of the data mining techniques within the intrusion detection context can effectively improve the detection accuracy, the detection speed, and enhance the system's own security [2]. Thus, as an intelligent analysis task, the AnomalyAgent provides the crossroads of multi-agents systems with the clustering technique, in particular the $\mathcal{AD}$-$\mathcal{C}$lust algorithm. The idea behind this technique is that the amount of normal connection data is usually overwhelmingly larger than that of intrusions [5]. Whenever this assumption holds, the anomalies and attacks can be detected based on cluster sizes, *i.e*, large clusters correspond to normal data, and the rest of the data points, which are outliers, correspond to attacks [19].

$\mathcal{AD}$-$\mathcal{C}$lust, (*Anomaly Detection-based Clustering*), is an unsupervised clustering algorithm introduced by Brahmi et *al.* in [4,5], to improve the quality of the K-means algorithm applied within the intrusion detection context. Indeed, the latter suffers from a greater time complexity, which becomes an extremely important factor within intrusion detection due to the very large packets sizes [15]. Moreover, the *number of clusters dependency* and the *degeneracy* constitute the drawbacks that hamper the use of K-means for anomaly detection [15]. In this respect, the $\mathcal{AD}$-$\mathcal{C}$lust algorithm combines two prominent categories of clustering, namely: distance-based [19] as well as density-based [8]. It exploits the advantages of the one to palliate the limitations of the other and vice versa.

The processing steps of our algorithm $\mathcal{AD}$-$\mathcal{C}$lust can be summarized as follows [4]:

1. Extraction of the density-based clusters that are considered as candidate initial cluster centers. The density-based clustering is used as a preprocessing step for the $\mathcal{AD}$-$\mathcal{C}$lust algorithm;

2. Compute the Euclidean distance between the candidate cluster center and the instance that will be assigned to the closest cluster. For an instance x_i and a cluster center z_i, the Euclidean distance is defined as:

$$distance(x_i, z_i) = \sqrt{\sum_{i=1}^{n} (x_i - z_i)^2} \quad (1)$$

3. The size of a neighborhood of instances is specified by an input parameter. We use the k' parameter to distinguish it from the k parameter used by the K-MEANS algorithm. Hence, k' specifies the minimal number of instances in a neighborhood and controls the granularity of the final clusters of the clustering-based density. If k' is set to a large value, then a few large clusters are found. To reduce the number of candidate clusters k' to the expected number k, we can iteratively merge the two most similar clusters. Otherwise, if k' is set too small, then many small clusters will be generated. The clusters will be split, new clusters will be created to replace the empty ones and the instances will be re-assigned to existing centers. This iteration will continue until there is no empty cluster. Consequently, the outliers of clusters will be removed to form new clusters, in which instances are more similar to each other. In this way, the value of initial cluster centers k will be determined automatically by splitting or merging clusters;
4. Within the detecting phase, the $\mathcal{AD}$-$\mathcal{C}$LUST algorithm performs the detection of intrusions. Thus, for each novel instance I the algorithm proceeds as follows:
   (a) Compute the Euclidean distance and find the cluster that presents the shortest distance with respect to I.
   (b) Classify I by the category of the closest cluster. Clearly, if the distance between I and the cluster of "normal" instances is the shortest one, then I will be a normal instance. Otherwise, I is an intrusion.

4 Experimental Results

In order to assess the overall performance of OCMAS-IDS in a realistic scenario, a prototype of the proposed architecture was implemented using Sun's Java Development Kit 1.4.1, the well known platform JADE[3] 3.7, the Eclipse and the JPCAP[4] 0.7. The ontology ASO is designed using PROTÉGÉ[5].

Through the carried out experiments, we have to stress on evaluating the performance of our system in terms of (*i*) the scalability-related criteria such as network bandwidth, detection delay and system response time; and (*ii*) the detection ability. During the evaluations, we compare the results of the OCMAS-IDS system *vs.* that of the centralized IDS SNORT [17] and the multi-agents

[3] Available at: http://jade.tilab.com
[4] Available at: http://netresearch.ics.uci.edu/kfujii/jpcap/doc/
[5] Available at: http://protege.stanford.edu/download/download.html

based ontology one MONI[6] [7]. All experiments were carried out on equivalent machines equipped with a 3GHz Pentium IV and 8GB of main memory. We used machines that were connected via a switch, thus forming a switched network. Moreover, we simulated attacks using the well known tool *Metasploit*[7] version 3.5.1. The simulated eight different attack types are:

- **attack1**: DoS Smurf;
- **attack2**: Backdoor Back Office;
- **attack3**: SPYWARE-PUT Hijacker;
- **attack4**: Nmap TCP Scan;
- **attack5**: Finger User;
- **attack6**: RPC Linux Statd Overflow;
- **attack7**: DNS Zone Transfer; and
- **attack8**: HTTP IIS Unicode.

4.1 The Scalability Evaluation

In order to test the scalability of OCMAS-IDS, we study the relationship between the bandwidth consumption and a number of attack types. Moreover, the variation of the detection delay according to the number of packets is evaluated. Additionally, we assess how the response time varies with respect to eight attack types.

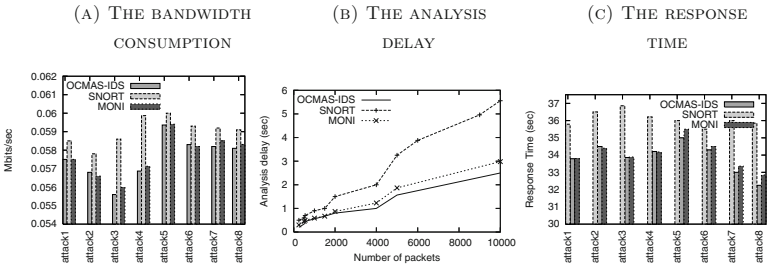


Fig. 3. The bandwidth consumption, the analysis delay and the response time of OCMAS-IDS *vs.* Snort and MONI

As depicted in Figure 3 (a), the maximum bandwidth consumed by OCMAS-IDS and MONI is lower compared to that of Snort. For example, the maximum bandwidth consumed by OCMAS-IDS is 0.06 Mbits/sec, which is very low as well. The reduction of the network bandwidth consumption is owe to the use of

[6] We thank Mrs. Djotio et *al.* [7] for providing us with the implementation of MONI system.

[7] Available at: http://www.metasploit.com/

the multi-agents system. Thus, the OCMAS-IDS system is not greedy in bandwidth consumption, which is definitely a desirable feature for any distributed system [4].

Besides, Figure 3(b) plots the detection delay against the number of packets, using the OCMAS-IDS, MONI and SNORT systems. According to this figure, we can answer the question: why the realization of the multi-agents IDS is advantageous? Clearly, the results show that the detection delay of both systems linearly increases with the number of packets. Moreover, the gap between both curves related to the detection delay of OCMAS-IDS and MONI is small, since both systems are based on multi-agents technology. In addition, Figure 3(b) highlights that our proposed system OCMAS-IDS is faster than the system SNORT. This can be explained by the fact that agents operate directly on the host whenever an action has to be taken, their response is faster than systems where actions were taken by the central controller, *i.e.*, SNORT.

Figure 3 (c) illustrates the response time required by OCMAS-IDS with respect to the attack types. On the one hand, we remark that the detection of all attack types, on average, result in lower response time compared to that of SNORT, due to its centralized detection engine. In addition, this figure proved how fast our system respond. For example, the response time of OCMAS-IDS was 35 seconds for attack5, which is absolutely negligible.

On the other hand, within MONI, the ontology model is developed under JADE. Differently, the ontology ASO of our system OCMAS-IDS is designed under PROTÉGÉ and queried with SWRL. The response time of OCMAS-IDS is better than that of MONI. The main reason is that in the case of OCMAS-IDS, the inferred model is computed only once before the matching starts and used throughout all the queries. Thus, the figure indicates that OCMAS-IDS outperforms MONI and permits the exploitation of the semantics of ASO.

To sum up, it is clear from the obtained results that the performance of the OCMAS-IDS will not deteriorate too much with the increase in the number of attacks, which is justified by its low bandwidth consumption, reduced detection delay and quick response time. Likewise, in case of more machines are connected to the network, the OCMAS-IDS system still withstand the load and swiftly deliver the results.

4.2 The Detection Ability

In order to evaluate the detection ability of an IDS, two interesting metrics are usually of use [4]: the *Detection Rate* (DR) and the *False Positive Rate* (FPR). Indeed, the DR is the number of correctly detected intrusions. On the contrary, the FPR is the total number of normal instances that were "incorrectly" considered as attacks. In this respect, the value of the DR is expected to be as large as possible, while the value of the FPR is expected to be as small as possible.

With respect to Figure 4 (a), we can remark that the FPR of OCMAS-IDS and MONI is significantly lower compared to that of SNORT. This fact is due to the adaptive mechanisms used by the agents, enabling both systems,

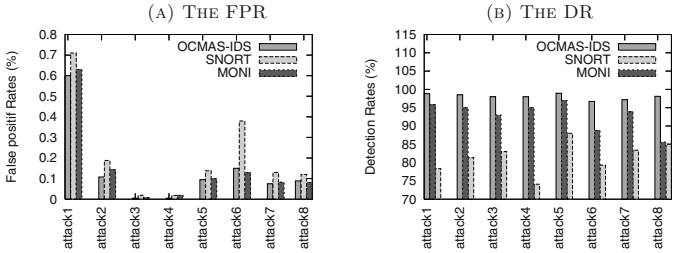


Fig. 4. The FPR and the DR of OCMAS-IDS *vs.* Snort and MONI

i.e., OCMAS-IDS and MONI, to better suit the environment. Consequently, the false alarms can be reduced correspondingly. For example, for attack3 the FPR of Snort can reach values as high as 0.019% compared to 0.007% of MONI and 0.005% of OCMAS-IDS.

Moreover, Figure 4 (b) shows that the DR of OCMAS-IDS is higher than that of MONI. Moreover, among the three investigated IDS, Snort has the lowest DR. For instance, for attack3, whenever OCMAS-IDS and MONI have the DR 97.9% and 94.9%, respectively, Snort has 74.1% DR. This is due to his centralized architecture.

Knowing that a main challenge of existing IDSs is to decrease the false alarm rates [4], the main benefit of our system is to lower the false alarm rate, while maintaining a good detection rate.

5 Conclusion

In this paper, we focused on a distributed architecture and multi-agents analysis of intrusions detection system to tackle the mentioned above challenges, *i.e.*, the high detection delay, the high bandwidth consumption as well as the low detection ability. Thus, we introduced a multi-agents intrusions detection system called $OCMAS\text{-}IDS$ based on an efficient ontology model, called ASO, as well as a clustering algorithm called $\mathcal{AD}$-$\mathcal{C}$lust. The carried out experimental results showed the effectiveness of the OCMAS-IDS system and highlighted that our system outperforms the pioneering systems fitting in the same trend.

Future issues for the present work mainly concern: (*i*) the alert correlation techniques by using the multi-agents system and ontology [12].

References

1. Abdoli, F., Kahani, M.: Ontology-based Distributed Intrusion Detection System. In: Proceedings of the 14th International CSI Computer Conference CSICC 2009, Tehran, Iran, pp. 65–70 (2009)
2. Azad, C., Jha, V.K.: Data Mining in Intrusion Detection: A Comparative Study of Methods, Types and Data Sets. International Journal of Information Technology and Computer Science (IJITCS) 5(8), 75–90 (2013)

3. Azevedoln, R.R., Dantas, E.R.G., Santos, R.C., Rodrigues, C., Almeida, M.J.S.C., Freitas, F., Veras, W.C.: An Autonomic Ontology-Based Multiagent System for Intrusion Detection in Computing Environments. The International Journal for Infonomics 3(1), 1–7 (2010)
4. Brahmi, I., Ben Yahia, S., Aouadi, H., Poncelet, P.: Towards a Multiagent-Based Distributed Intrusion Detection System Using Data Mining Approaches. In: Cao, L., Bazzan, A.L.C., Symeonidis, A.L., Gorodetsky, V.I., Weiss, G., Yu, P.S. (eds.) ADMI 2011. LNCS, vol. 7103, pp. 173–194. Springer, Heidelberg (2012)
5. Brahmi, I., Ben Yahia, S., Poncelet, P.: $\mathcal{AD}$-$\mathcal{C}$LUST: Détection des anomalies basée sur le Clustering. In: Atelier Clustering Incrémental et Méthodes de Détection de Nouveauté en conjonction avec 11ème Conférence Francophone d'Extraction et de Gestion de Connaissances EGC 2011, Brest, France, pp. 27–41 (2011)
6. Brahmkstri, K., Thomas, D., Sawant, S.T., Jadhav, A., Kshirsagar, D.D.: Ontology Based Multi-Agent Intrusion Detection System for Web Service Attacks Using Self Learning. In: Meghanathan, N., Nagamalai, D., Rajasekaran, S. (eds.) Networks and Communications (NetCom2013). LNEE, vol. 284, pp. 265–274. Springer, Heidelberg (2014)
7. Djotio, T.N., Tangha, C., Tchangoue, F.N., Batchakui, B.: MONI: Mobile Agents Ontology based for Network Intrusions Management. International Journal of Advanced Media and Communication 2(3), 288–307 (2008)
8. Duan, L.: Density-Based Clustering and Anomaly Detection. In: Mircea, M. (ed.) Business Intelligence - Solution for Business Development, pp. 79–96 (2012)
9. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML (2004), http://www.w3.org/Submission/SWRL/
10. Isaza, G.A., Castillo, A.G., López, M., Castillo, L.F.: Towards Ontology-Based Intelligent Model for Intrusion Detection and Prevention. Journal of Information Assurance and Security 5, 376–383 (2010)
11. Kim, G., Lee, S., Kim, S.: A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection With Misuse Detection. Expert Systems with Applications, 41(4, pt. 2), 1690–1700 (2014)
12. Li, W., Tian, S.: An Ontology-Based Intrusion Alerts Correlation System. Expert Systems with Applications 37(2010), 7138–7146 (2010)
13. Mandujano, S., Galvan, A., Nolazco, J.A.: An Ontology-Based Multiagent Approach to Outbound Intrusion Detection. In: Proceedings of the International Conference on Computer Systems and Applications, AICCSA 2005, Cairo, Egypt, pp. 94–I (2005)
14. PinzóN, C.I., De Paz, J.F., Herrero, Á., Corchado, E., Bajo, J., Corchado, J.M.: idMAS-SQL: Intrusion Detection Based on MAS to Detect and Block SQL Injection Through Data Mining. Information Sciences 231, 15–31 (2013)
15. Ranjan, R., Sahoo, G.: A New Clustering Approach For Anomaly Intrusion Detection. International Journal of Data Mining and Knowledge Management Process (IJDKP) 4(2), 29–38 (2014)
16. Raskin, V., Hempelmann, C.F., Triezenberg, K.E., Nirenburg, S.: Ontology in Information Security: A Useful Theoretical Foundation and Methodological Tool. In: Proceedings of the 2001 Workshop on New Security Paradigms, NSPW 2001, Cloudcroft, New Mexico, pp. 53–59 (2001)
17. Roesch, M.: Snort - Lightweight Intrusion Detection System for Networks. In: Proceedings of of the 13th USENIX Conference on System Administration (LISA 1999), Seattle, Washington, pp. 229–238 (1999)

18. Sodiya, A., Ojesanmi, O., Akinola, O.C., Aborisade, O.: Neural Network based Intrusion Detection Systems. International Journal of Computer Applications 106(18), 19–24 (2014)
19. Syarif, I., Prugel-Bennett, A., Wills, G.: Unsupervised Clustering Approach for Network Anomaly Detection. In: Proceedings of the 4th International Conference on Networked Digital Technologies (NDT 2012), Dubai, AE, pp. 135–145 (2012)
20. Undercoffer, J., Joshi, A., Pinkston, J.: Modeling Computer Attacks: An Ontology for Intrusion Detection. In: Proceedings of the 6th International Workshop on the Recent Advances in Intrusion Detection, Pittsburgh, PA, USA, pp. 113–135 (2003)

On Copulas-Based Classification Method for Intrusion Detection

Abdelkader Khobzaoui[1(✉)], Mhamed Mesfioui[2], Abderrahmane Yousfate[3], and Boucif Amar Bensaber[2]

[1] Computer sciences Department, University of Saida
[2] Département de mathématiques et informatique, Université du Québec, Trois-Rivières, C.P, 500, Québec, Canada, G9A 5H7
[3] Laboratoire de mathématiques (LDM), University of Sidi Bel Abbès

Abstract. The intent of this paper is to develop a nonparametric classification method using copulas to estimate the conditional probability for an element to be a member of a connected class while taking into account the dependence of the attributes of this element. This technique is suitable for different types of data, even those whose probability distribution is not Gaussian. To improve the effectiveness of the method, we apply it to a problem of network intrusion detection where prior classes are topologically connected.

Keywords: Intrusion detection · Classification · Copula function · Copula density estimator · Empirical copula

1 Introduction

Let a set of d attributes $(a_1, a_2, \ldots, a_d)$ characterizing a vectorial space E. Let also $(x_1, x_2, \ldots, x_n)$ a set of E used as a learning set over m classes denoted $(\omega_1, \omega_2, \ldots, \omega_m)$ which are actually some disjoint subsets of E. To avoid the use of some predetermined probability laws of the attributes systematically, we intent to build a copulas-based classification model that estimates the true attributes laws and their dependency. Then one assigns each entity of E to its most likely class ω_i; $i \in \{1, \ldots, m\}$. This entity must be well-assigned when it verifies an optimal probabilistic criterion.

In deterministic classification, this model builds, over the set E, an equivalence relation $\mathcal{R} \subset E \times E$ where $E/\mathcal{R}$ is a partition of E. In nondeterministic classification, for some adapted risks, classes are built using probability distributions. Each realization of the observed phenomenon distributes all elements over the different classes which yields to a partition of E. Partition changes with realizations (samples). To assign k elements o7ver m classes, in the deterministic case, one has only k steps to carry out all the affectations; each step requires m simplified tests. However, in the nondeterministic case, if one enumerates all possibilities for distributing k entities over m classes, then one finds m^k possibilities; each possibility requires k assigning steps. Each element is assigned to

Funded in part by DGRSDT, Algiers (PNR : Data mining and applications)

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 394–405, 2015.
DOI: 10.1007/978-3-319-19578-0_32

the class ω_j via a conditional probability $f(x \mid j)$ which can be estimated using the training data. Actually, we seek the most likely class k (maximum likelihood estimation) solution of : $k = \arg\max_j (f(x \mid j))$ where $f(x \mid j)$ denotes the conditional probability density function for x being a member of group ω_j.

To reduce the complexity of the problem, one assigns elements to their respective classes as in deterministic affectation. Elements whose ranges are near apexes are the most likely affected. In this case, each step requires m complicated tests; that means $k.m$ complicated tests.

In the following we'll denote $f^j(x)$ instead $f(x \mid j)$.

Many applications algorithms and models have been proposed to estimate this conditional probability density function : kernel-density estimator [32], k-nearest-neighbours (KNN) method [19], Learning Vector Quantisation (LVQ) [12], Support Vector Machines (SVM)[20] ...

In this work we present the use of the empirical copula function as an alternative for modeling dependence structure in a supervised probabilistic classifier. The set E is identified to a vector space $\mathbf{R}^d$ over the field $\mathbf{R}$ and we use the law of the considered phenomenon over E which can be well estimated if learning sample is sizable. So, the conditional probability density function $f^j(x)$ is estimated according the following algorithm:

Algorithm 1. Conditional probability density estimation

Require: :

- $\{X_i\}_{i=1}^n$ an iid random sample from a $d-$dimensional distribution F with density f.
- $\Omega = \{\omega_1, \cdots, \omega_m\}$ m learning classes.

1. **for** each $j \in \{1, \cdots, m\}$ **do**
2. Transform the observations X_i^j to $U_i^j = F_{ni}^j(X_i)$ where F_{ni}^j estimates the ith marginal distribution restricted to a class ω_j and X_i^j denotes observation from the class ω_j
3. Estimate the marginal densities f_i^j for class ω_j.
4. Estimate the joint density of the transformed data restricted to the class ω_j. this density w'il be noted c^j and it is equivalent to the copula density.
5. Estimate the joint density of the original data restricted to a class ω_j by:

$$f^j(x) = c^j\left(F_1(x_1), \ldots, F_d(x_d)\right) \prod_{i=1}^{d} f_i^j(x_i)$$

6. **end for**

This approach allows to mitigate the curse of dimensionality and to treat the data in all situations even if the variance does not exist. It considers also the non-linear relationships between attributes.

New observation x will be affected to the class ω_r such that

$$r = \arg\max_j f^j(\mathbf{x})$$

The content of the paper is the following: The second section of the paper gives a short mathematical background of copula functions, Section 3 presents a copula based probabilistic model for classification. Section 4 presents the experimental setting to detect and identify intrusion in computer network and Section 5 summarizes the conclusions

2 Copulas Theory

Copulas play an important role in several areas of statistics and in Machine Learning as a tool of studying scale-free measures of dependence and as starting point for constructing families of bivariate distribution especially in applications where nonlinear dependencies are common and need to be represented.

The best definition of a copula is that given by referring to well know Sklar's theorem [28], [18], which states how a copula function is related to joint distribution functions.

Theorem 1 (Sklar's Theorem). *Let F be any d-dimensional distribution function over real-valued random variables with marginals $f_1, f_2, ..., f_d$,then there exists a copula function C such that for all $x \in \bar{\boldsymbol{R}}^d$*

$$F\left(x_1, \ldots, x_d\right) = C(f_1(x_1), \ldots, f_d(x_d)) \tag{1}$$

where $\bar{\boldsymbol{R}}$ denotes the extended real line $[-\infty, \infty]$ and $C : [0,1]^p \to [0,1]$.

The copula distribution can also be stated as joint distribution function of standard uniform random variables:

$$C(u_1, \ldots, u_p) = P\left(U_1 \leq u_1, \ldots, U_p \leq u_p\right) \tag{2}$$

where $U_i \sim U(0,1)$ for $i = 1, \ldots, p$.

Note that if $f_1(x_1), \ldots, f_d(x_d)$ in (1) are all continuous, then C is unique. Otherwise, C is uniquely determined on $Ran(f_1) \times Ran(f_2) \times \cdots \times Ran(f_d)$, where Ran stands for the range.

Conversely, if C is an d-copula and $f_1, \ldots, f_d$ are distribution functions, then the function F defined above is an d-dimensional distribution function with margins $f_1, \ldots, f_d$. For the proof, see [28].

From Sklar's theorem we see that for continuous multivariate distribution functions, the univariate margins and the multivariate dependence structure can be separated, and the dependence structure can be represented by a copula.

An important consequence of theorem 1 is that the d-dimensional joint density F and the marginal densities $f_1, f_2, \ldots, f_d$ are also related:

$$f(x_1, \ldots, x_d) = c\left(F_1(x_1), \ldots, F_d(x_d)\right) \prod_{i=1}^{d} f_i(x_i) \tag{3}$$

where c denotes the density of the copula C. The equation (3) shows that the product of marginal densities and a copula density builds a d-dimensional joint density.

The unique copula function related to the multivariate distributions F with continues margins $f_i; 1 \leq i \leq d$ is determined by

$$C(u_1, \ldots, u_d) = F(F_i^{-1}(u_1), \ldots, F_i^{-1}(u_d)) \tag{4}$$

where

$$F_i^{-1}(s) = \{t \mid F_i(t) \geq s\} \tag{5}$$

denote the pseudo-inverse of the univariate margins $F_1, \cdots, F_d$.
Copulas are essentially a way of transforming the random variable $(X_1, \cdots, X_d)$ into another random variable $(U_1, \cdots, U_d) = (F_1(X_1), \cdots, F_d(X_d))$ having the margins uniform on $[0, 1]$ and preserving the dependence among the components. Without the continuity assumption, care must be taken to use equation (4); see [21] or [17].

3 Copula Function Estimation

To estimate copula functions, the first issue consists in specifying how to estimate separately the margins and the joint law. Moreover, some of these functions can be fully known. Depending on the assumptions made, some quantities have to be estimated parametrically, or semi or even non-parametrically. In the latter case, we have to choose between the usual methodology of using "empirical counterparts" and invoking smoothing methods well-known in statistics: kernels, wavelets, orthogonal polynomials, nearest neighbors,... A non-parametric estimation of copula treats both the copula and the marginals parameter-free and thus offers the greatest generality.

Unlike the marginal and the joint distributions which are directly observable, a copula is a hidden dependence structure. This makes the task of proposing a suitable parametric copula model non-trivial and is where a non-parametric estimator can play a significant role.

Indeed, a non-parametric copula estimator can provide initial information needed in revealing and subsequent formulation of an underlying parametric copula model[3].

Non-parametric estimation of copulas dates back to Deheuvels [6], who proposed the so-called empirical copula defined by

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(F_{n,1}(X_{i1}) \leq u_1, \ldots, F_{n,d}(X_{i,d}) \leq u_d\right) \tag{6}$$

where $F_{n,i}$ are the empirical distribution function given by

$$F_{n,j}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{i,j} \leq x) \tag{7}$$

with j=1,...,d and $\mathbf{u} \in [0, 1]^d$.

Let R_i be the rank of X_i among the sample $X_1, \ldots, X_n$. Observe that C_n is a function of ranks $R_1, R_2, \ldots, R_n$, because $F_{n,j}(X_i) = \dfrac{R_{i,j}}{n} \quad i = 1, \ldots, n$, namely;

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\frac{R_{i,1}}{n} \leq u_1, \ldots, \frac{R_{i,d}}{n} \leq u_d \right). \tag{8}$$

From this representation, one can consider $C_n(\mathbf{u})$ as discrete multivariate distribution with uniform marginals takings values in the set $\left[\frac{1}{n}, \frac{2}{n}, \ldots, 1\right]$. and so his density:

$$c_n(\mathbf{u}) = \frac{\partial C(u_1, ..., u_d)}{\partial u_1, \cdots, \partial u_d} \tag{9}$$

can be estimated by a standard kernel function:

$$\hat{c}_n(\mathbf{u}) = \frac{1}{n} \sum_{j=1}^{n} \prod_{i=1}^{d} h_i^{-1} K\left(\frac{u_i - U_{ji}}{h_i^{-1}} \right) \tag{10}$$

where U_i is the transformed of the original data given by $U_i = F_{n,i}^{j}(X_i)$ as described above. And a uni-variate kernel function $K(u)$ is any functions satisfying the following conditions:

(a) $K(x) \geq 0 \quad and \quad \int_{\mathbf{R}} K(x)dx = 1$
(b) $\int_{\mathbf{R}} xK(x)dx = 0$ (Symmetric about the origin)
(c) Has finite second moment e.g. $\int_{\mathbf{R}} x^2 K(x)dx < \infty$

So, we have to choice both kernel function K and their smoothing parameter or bandwidth h. Actually, selection of K is a problem of less importance, and different functions that produce good results can be used (see table 1 for some examples).

In this paper, we use the Gaussian one given by:

$$K(v) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{v}{2}).$$

In practice, the choice of an efficient method for the calculation of h; for an observed data sample is a more complex problem, because of the effect of the bandwidth on the shape of the corresponding estimator. If the bandwidth is small, we will obtain an under-smoothed estimator, with high variability. On the contrary, if the value of h is big, the resulting estimator will be very smooth and farther from the function that we are trying to estimate[23](see figure 1).

For evaluating the tradeoff between bias and variance. Silverman[31] has suggested a frequently used rule-of-thumb bandwidth

$$h_n = 0.9(min(\hat{\sigma}, \frac{IQR}{1.34})n^{\frac{1}{5}},$$

Table 1. some kernel functions

	Kernel	K(x)
1	uniform	$\frac{1}{2}\mathbf{1}_{(\lvert x\rvert\leq 1)}$
2	Epanechnikov	$\frac{3}{4}(1-x^2)\mathbf{1}_{(\lvert x\rvert\leq 1)}$
3	Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x}{2}\right)$
4	triangular	$(1-\lvert x\rvert)\mathbf{1}_{(\lvert x\rvert\leq 1)}$
5	Triweight	$\frac{35}{32}(1-x^2)^3\mathbf{1}_{(\lvert x\rvert\leq 1)}$
6	Tricube	$\frac{70}{81}(1-x^3)^3\mathbf{1}_{(\lvert x\rvert\leq 1)}$
7	Biweight(Quartic)	$\frac{15}{16}(1-x^2)^2\mathbf{1}_{(\lvert x\rvert\leq 1)}$
8	Cosine	$\frac{\pi}{4}cos(\frac{\pi}{2}x)\mathbf{1}_{(\lvert x\rvert\leq 1)}$

where IQR is the interquartile range (the difference between the 75th and 25th percentile) and $\hat{\sigma}$ is the sample standard deviation. Like all desirable bandwidth selection procedures, this bandwidth gets smaller as the number of observations n increases, but does not go to zero "too fast"[8].

4 The Probabilistic Classifier

As noted, the aim of this work is to develop a non-parametric classification method using a copula functions to estimate the conditional probability density $f^j(x)$ for one element x being a member of class ω_j. Actually, we use the empirical copula function estimator as tool to estimating $f^j(x)$ given by the equation 3.

Consider a set of m class $\omega_1, \ldots, \omega_m$. Each class ω_j is characterized by a d-random vector $\mathbf{X}^j = (X_1^j, \ldots, X_d^j)$. Let $(X_{11}^j, \ldots, X_{1d}^j), \ldots, (X_{n1}^j, \ldots, X_{nd}^j)$ be a random sample arises from the class ω_j. The distribution of component $\mathbf{X}_i^j$ of the random vector $\mathbf{X}^j$ may be estimated by

$$F_{n,i}^j(x_i) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{I}\left(X_{ki}^j \leq x_i\right).$$

The density function of this component is also estimated by

$$\hat{f}_i^j(x_i) = \frac{1}{n}\sum_{j=1}^{n}K(x_i - X_{ji})$$

where

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$$

The density function of the random vector $\mathbf{X}_j$ can be estimated by

$$\hat{f}^j(\mathbf{x}) = \hat{c}^j\left(F_{n,1}^j(x_1), \ldots, F_{n,d}^j(x_d)\right)\prod_{i=1}^{d}\hat{f}_i^j(x_i) \tag{11}$$

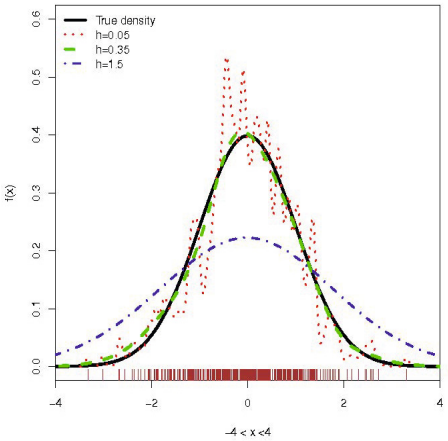


Fig. 1. Standard normal distribution density and its kernel density estimate (KDE) with different bandwidths obtained from a random sample of size 500. Solid line : True density (standard normal). Dotted line : KDE with h=0.05. Dashed line : KDE with h=0.35. Dot-dash line : KDE with h=1.5

where $\hat{c}^j$ denotes the estimator of the copula density associated to a random vector $\mathbf{X}^j$ estimated by a standard kernel function as described in equation 10.

So, all elements of our classifier are constructed, namely: $\hat{c}$ the copula density estimators, $\hat{f}_i^j$ the marginal density estimators, and $\hat{f}^j$ the joint density estimators.

The goal of the classifier is to determine, given a new observation x, its most likely corresponding class ω_r which is chosen as follow:

$$r = \arg\max_j \hat{f}^j(\mathbf{x})$$

Finally, we will describe the main steps of our classifier:

Algorithm 2. The probabilistic classifier algorithm

1. Let $\mathbf{x} = (x_1, \ldots, x_d)$ a new observation.
2. For each $j \in \{1, \cdots, m\}$ Do
3. For each $i \in \{1, \cdots, d\}$ Do
   - $u_i^j \leftarrow F_{n,i}^j(x_i)$
   - Compute $\hat{f}_i^j(x_i)$
4. EndFor
5. Compute $\hat{c}^j \left(F_{n,1}^j(x_1), \ldots, F_{n,d}^j(x_d)\right)$ as described above
6. Compute $\hat{f}^j(\mathbf{x})$ from equation(11)
7. EndFor
8. affect the observation x to the class ω_r such that

$$r = \arg\max_j \hat{f}^j(\mathbf{x})$$

5 Application

To verify the effectiveness and the feasibility of the proposed algorithm, we use the KDD'99 dataset ([5]), was originally provided by MIT Lincoln, Labs which contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a real-world military network environment.

The KDD'99 dataset includes a set of 41 features, gathered in 7 symbolic ones and 34 numeric. A complete description of all 41 features is available in [5]. These features are divided into four categories:

1. The intrinsic features of a connection, which includes the basic features of individual TCP connections. For example, duration of the connection, the type of the protocol (tcp, udp, etc), network service (http, telnet, etc), etc.
2. The content feature within a connection suggested by domain knowledge is used to assess the payload of the original TCP packets, such as number of failed login attempts.
3. The same host features examine established connections in the past two seconds that have the same destination host as the current connection, and calculate statistics related to the protocol behavior, service, etc.
4. The similar same service features examine the connections in the past two seconds that have the same service as the current connection.

These features describe 23 behaviors of which one corresponds to a normal traffic and the 22 others correspond to attacks which are gathered in four categories as summarized in table 2 :

1. DOS (Denial of service): making some computing or memory resources too busy so that they deny legitimate users access to these resources.
2. R2L (Root to local): unauthorized access from a remote machine according to exploit machine's vulnerabilities.
3. U2R (User to root): unauthorized access to local super user (root) privileges using system's susceptibility.
4. PROBE: host and port scans as precursors to other attacks. An attacker scans a network to gather information or find known vulnerabilities.

We used the train data-set which is about 494 020 connection record and test data-set is about 4 898 431. First, the symbolic variables are converted to numeric ones, the zero colones and repeated rows are removed we obtained 145 586 rows for training and 1 074 992 for test.

Calculations are performed under the R Environment for Statistical Computing [24] [25] using the parallel packages snow[29] and snowfall[30] under Linux RedHat enterprise 6 workstation on Intel Core I7 with 16 Go of Ram and 4 physical cores.

As confusion matrix between all behaviors is too big, we present in Table 3 table a summarized confusion matrix between the five categories of behaviors(described above). This condensed representation allows us to compare our results with those presented by other authors which have used the same data set.

Table 2. Class label in KDD '99 Dataset

Id-Attack	Attack	Category
1	back	dos
2	buffer_overflow	u2r
3	ftp_write	r2l
4	guess_passwd	r2l
5	imap	r2l
6	ipsweep	probe
7	land	dos
8	loadmodule	u2r
9	multihop	r2l
10	neptune	dos
11	nmap	probe
12	normal	normal
13	perl	u2r
14	phf	r2l
15	pod	dos
16	portsweep	probe
17	rootkit	u2r
18	satan	probe
19	smurf	dos
20	spy	r2l
21	teardrop	dos
22	warezclient	r2l
23	warezmaster	r2l

Conditional distributions are on rows. For example the first row means that normal behavior is identified as normal with estimate probability 97.375% (True Negative Attacks). It is identified as DOS behavior with estimate probability 0.406%, as PROB behavior with estimate probability 2.038% as as R2L behavior with estimate probability 0.175% and U2R behavior with estimate probability 0.006%. These four last identifications are said "False Positive Attacks". From second to fifth rows when behavior is identified as Normal, this identification is said "False Negative Attacks" else it is said "True Positive Attacks".

In order to evaluate the performances of our method, we compare the our results with those obtained by other authors which have used the same data set.

Table 3. Results by Attacks categories

	Normal	Dos	Probe	R2L	U2R
Normal	97.375	0.406	2.038	0.175	0.006
Dos	0.068	97.357	2.563	0.010	0.002
Probe	4.928	4.199	90.548	0.094	0.231
R2L	0.000	0.000	0.000	100.000	0.000
U2R	0.000	0.000	0.000	0.000	100.000

Table 4. Performance comparison of proposed Algorithm

Method	Normal	Dos	Probe	U2R	R2L
MCAD[26]	95.20	99.20	97.0	72.80	69.20
KDD cup 99 Winer [22]	99.50	97.10	83.30	13.20	08.40
GP Multi- Transformation[10]	99.93	98.81	97.29	45.20	80.22
C.N.B.D.[11]	99.72	99.75	99.25	99.20	99.26
PNRule[1]	99.50	96.9	73.20	06.60	10.70
ESC-IDS-1[33]	98.20	99.5	84.10	14.10	31.50
Prazen-window N.I.D.[34]	97.38	96.71	99.17	93.57	31.17
Model 1(a)[13]		97.40	83.80	32.80	10.70
SVM-IDS [9]	99.80	92.5	98.30	05.10	70.20
NN Classifier wiht GDA[27]	98.95	98.63	96.50	24.12	12.08
SVM+DGSOT[14]	95.00	97.00	91.00	23.00	43.00
I.C.A.[7]	69.60	98.00	100.00	71.40	99.20
C.L.C. [15]	73.95	99.88	87.83	61.36	98.50
Multi- PD[16]		97.30	88.70	29.80	09.60
ADWICE[4]		98.30	96.00	81.10	70.80
Our method	**97.375**	**97.357**	**90.548**	**100.00**	**100.00**

6 Conclusion

The method proposed, in this paper, presents many interesting advantages with respect to previous proposals in the field of intrusion detection, when applied to KDDCup'99 data set.

The obtained results, confirm the fact that copulas are flexible and powerful tool of studying scale-free measures of dependence and as starting point for constructing families of multivariate distribution especially in applications where nonlinear dependencies involved in the study and need to be represented. That occurs essentially when attributes probability laws are non-gaussian.

References

1. Agarwal, R., Joshi, M.V.: PNrule: A New Framework for Learning Classifier Models in Data Mining. In: Proceedings of the First SIAM International Conference on Data Mining, Chicago, IL, USA, April 5-7 (2001)
2. Chao, M., Xin, S.Z., Min, L.S.: Neural network ensembles based on copula methods and Distributed Multiobjective Central Force Optimization algorithm. Engineering Applications of Artificial Intelligence 32, 203–212 (2014)
3. Chen, S.X., Huang, T.-M.: Nonparametric estimation of copula functions for dependence modelling. The Canadian Journal of Statistics 35(2) (2007)
4. Burbeck, K., Nadjm-Tehrani, S.: ADWICE – anomaly detection with real-time incremental clustering. In: Park, C.-S., Chee, S. (eds.) ICISC 2004. LNCS, vol. 3506, pp. 407–424. Springer, Heidelberg (2005)
5. DARPA Intrusion Detection Data set, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html

6. Deheuvels, P.: La fonction de dépendance empirique et ses propriétés. Un test non paramétrique díndépendance. Bulletin de la classe des sciences. Acadimie Royale de Belgique 65, 274–292 (1979)
7. Dayu, Y., Qi, H.: A Network Intrusion Detection Method using Independent Component Analysis. In: International Conference on Pattern Recognition (ICPR), Tampa, FL, pp. 8–11 (2008)
8. DiNardo, J., Tobias, J.L.: Nonparametric Density and Regression Estimation. Journal of Economic Perspectives 15(4), 11–28 (2001)
9. Eid, H.F., Darwish, A., Hassanien, A.E., Ajith, A.: Principle Components Analysis and Support Vector Machine based Intrusion Detection System. In: 10th International Conference on Intelligent Systems Design and Applications (2010)
10. Faraoun, K.M., Boukelif, A.: Securing network trafic using geneticaly evolved transformations. Malaysian Journal of Computer Science 19(1) (2006)
11. Farid, D.M., Harbi, N., Rahma, Z.M.: Combining naive bayes and decision tree for adaptive intrusion detection. International Journal of Network Security & Its Applications (IJNSA) 2(2) (2010)
12. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer (2000)
13. Nguyen, H.A., Choi, D.: Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. In: Ma, Y., Choi, D., Ata, S. (eds.) APNOMS 2008. LNCS, vol. 5297, pp. 399–408. Springer, Heidelberg (2008)
14. Khan, L., Awad, M., Thuraisingham, B.: A new intrusion detection system using support vector machines and hierarchical clustering. The International Journal on Very Large Data Bases 16(4), 507–521 (2007)
15. Levin, I.: KDD-99 Classifier Learning Contest LLSoft's Results Overview. ACM SIGKDD Explorations Newsletter 1(2), 67–75 (2000)
16. Maheshkumar, S., Gursel, S.: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. In: Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, Las Vegas (MLMTA 2003), vol. 1, pp. 209–215 (2003)
17. Marshall, A.: Copulas, marginals and joint distributions. In: Rüschendorff, L., Schweizer, B., Taylor, M. (eds.) Distributions with Fixed Marginals and Related Topics, pp. 213–222. Institute of Mathematical Statistics, Hayward (1996)
18. Mayor, G., Suñer, J., Torrens, J.: Sklar's theorem in finite settings. IEEE Transactions on Fuzzy Systems 15(3), 410–416 (2007)
19. Michie, D., Spiegelhalter, D.J., Tayler, C.C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood Series in Artificial Intelligence. Prentice Hall, Upper Saddle River (1994)
20. Müller, K.R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B.: An Introduction to Kernel-Based Learning Algorithms. IEEE Transactions on Neural Networks 12, 181–201 (2001)
21. Nelsen, R.: An Introduction to Copulas, 2nd edn. Springer, New York (2006)
22. Pfahringer, B.: Winning the KDD99 classification cup: bagged boosting. ACM SIGKDD Explorations Newsletter 1(2), 65-66 (2000)
23. Quintela-del-Río, A., Estévez-Pérez, G.: Nonparametric Kernel Distribution Function Estimator with kerdiest: An R Package for Bandwidth Choice and Applications. Journal of Statistical Software 50(8) (2012)
24. The Comprehensive R Archive Network, http://cran.r-project.org/
25. Rossiter, D.G.: Tutorial: Using the R Environment for Statistical Computing: An example with the Mercer & Hall wheat yield dataset. University of Twente, Faculty of Geo-Information Science & Earth Observation (ITC) Enschede, NL (2014)

26. Santosh, K., Sumit, K., Sukumar, N.: Multidensity Clustering Algorithm for Anomaly Detection Using KDD'99 Dataset. Advances in Computing and Communications 190(pt.8), 619–630 (2011)
27. Singh, S., Silakari, S.: Generalized Discriminant Analysis algorithm for feature reduction in Cyber Attack Detection System. International Journal of Computer Science and Information Security 6(1) (2009)
28. Sklar, A.: Fonction de répartition á n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231 (1959)
29. snow: Simple Network of Workstations, http://cran.r-project.org/web/packages/snow/index.html
30. snowfall: Easier cluster computing (based on snow), http://cran.r-project.org/web/packages/snowfall/index.html
31. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. Chapman and Hall, London (1986)
32. Terrell, D.G., Scott, D.W.: Variable kernel density estimation. Annals of Statistics 20(3), 1236–1265 (1992)
33. Toosi, A.N., Kahani, M.: A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers. Computer Communications 30, 2201–2212 (2007)
34. Yeung, D.Y., Chow, C.: Parzen-window Network Intrusion Detectors. In: 16th International Conference on Pattern Recognition, Quebec, Canada, pp. 11–15 (2002)

On-Off Attacks Mitigation against Trust Systems in Wireless Sensor Networks

Nabila Labraoui [1(✉)], Mourad Gueroui [2], and Larbi Sekhri[3]

[1] STIC Laboratory, University of Tlemcen, Tlemcen, Algeria
nabila.labraoui@mail.univ-tlemcen.dz
[2] PRISM Laboratory, University of Versailles Saint-Quentin en Yvelines, Versailles, France
mourad.guerroui@prism.uvsq.fr
[3] ICN Laboratory, University of Oran, Es Senia, Algeria
larbi.sekhri@univ-oran.dz

Abstract. Trust and reputation systems have been regarded as a powerful tool to defend against insider attacks caused by the captured nodes in wireless sensor networks (WSNs). However, trust systems are vulnerable to on-off attacks, in which malicious nodes can opportunistically behave good or bad, compromising the network with the hope that bad behavior will be undetected. Thus, malicious nodes can remain trusted while behaving badly. In this paper, we propose O^{2}Trust, On-Off attack mitigation for Trust systems in wireless sensor networks. O^{2}Trust adopts the penalty policy against the misbehavior history of each node in the network as a reliable factor that should influence on the calculation of the trust value. This punishment future helps to perceive malicious node that aim to launch intelligent attacks against trust-establishment and consequently on-off attack is mitigated efficiently.

1 Introduction

Wireless sensor networks (WSNs) [1] provide a technological basis for many different security critical applications such as critical infrastructure monitoring, healthcare and battlefield. However, WSNs are often deployed in unattended, harsh and hostile environment that makes them under the threat of various types of attacks, including node compromise. In a node capture attack, an adversary tries to physically tamper with a node in order to extract the cryptographic secrets. Hence, the compromised node can participate in the network as a legitimate node and cannot be identified whether it is genuine or not. This attack can give rise to many subsequent powerful insider attacks [2]. Unfortunately, traditional safety mechanisms based on cryptography, cannot adequately defend against network insider attacks, although they are effective to outsider attacks [3].

Trust and reputation systems have been regarded as a powerful tool to defend against insider attacks caused by the captured nodes in WSNs [4]. Generally, trust establishment is used to record feedback about the security evaluations of other nodes. Thus, efficient trust management systems can help well-behaved nodes to avoid working with misbehaving nodes, as well as to detect these malicious ones [5]. How-

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 406–415, 2015.
DOI: 10.1007/978-3-319-19578-0_33

ever, building a robust trust and reputation system presents several important challenges on its own [6], because it is susceptible to attacks such as bad-mouthing and on-off attacks [7, 8]. In this work we consider the on-off attack in which malicious nodes can opportunistically behave good or bad, compromising the network with the hope that bad behavior will be undetected. Malicious nodes can remain trusted while behaving badly. As it is mentioned in [8], almost all reputation-based trust models are vulnerable to on-off attack, because they focus more on recent behavior of the node rather than comprehensively combining the nodes' past behavior with its instantaneous behavior. As a consequence, a malicious node can easily dissimulate any misbehavior history by either displaying good behavior or waiting during later time periods to increase its trust value. By this way, it continues its attack.

To address the above problem, we present in this paper O^2Trust: On-Off attack mitigation for Trust systems in wireless sensor networks. O^2Trust adopts the *Penalty Policy* against the misbehavior history of each node in the network. Unlike previous trust models that focus on recent behavior and thus are not sensitive enough to perceive contradictory behavior, in our proposal, we focus on frequency misbehavior history as a reliable factor that should influence on the calculation of the trust value for a node. This punishment future helps to perceive malicious node that aim to launch intelligent attacks against trust-establishment and consequently on-off attack is mitigated.

The rest of this paper is organized as follows. In Section 2, we present an overview of related works. Section 3 describes the proposed trust model. Evaluation results and theoretical analyses of the proposed model are provided in Section 4 and Section 5. Section 6 concludes the paper.

2 Related Works

Ganeriwal and Srivastava [9] proposed the first reputation and trust based model designed and developed exclusively for sensor networks; the RFSN (Reputation-based Framework for high integrity Sensor Networks) model uses the Beta distribution as a mathematical tool to represent and continuously update trust and reputation. To differently weight the old and new interactions, an aging factor is introduced for trust updating; more weight is given to recent interactions. Chen proposed in [10], a Task-based Trust framework for Sensor Networks (TTSN), where sensor nodes maintain reputation for neighbor nodes of several different tasks and use the reputation to evaluate their trustworthiness. The method for trust calculation and trust updating is almost the same as described in RSFN [9]. Sheikh et al. [11] proposed GTMS a Group-based Trust Management Scheme, in which the whole group will get a single trust value. He et al. [12] proposed attack-resistant and lightweight trust management scheme (ReTrust) for medical sensor network followed a hierarchical architecture, comprised of master nodes and sensor nodes. The authors use the window mechanism to forget previous actions. Moreover, they introduce an aging-factor parameter, which is different for each time unit m in the window.

3 The Proposed Trust Model: O²Trust

In this section, we will present a novel trust model for wireless sensor networks named on-off attack mitigation for trust systems in WSNs (O²Trust).

3.1 Overview

The design of O²Trust is based on *penalty policy* that is based on misbehavior history. In O²Trust, the evaluation model reflects nodes' real-time trust state accurately and is very sensitive to past malicious actions. This policy deals efficiency with the dynamic and contradictory misbehavior of malicious nodes. Dynamicity of the misbehavior is not considered under traditional trust estimation models because trust values are obtained based on current behavior, which does not indicate continuity of misbehavior. In other terms, only weight of measured misbehavior is considered rather than periodicity of the misbehavior along with weight of measured misbehavior.

Unlike the previous trust models, the trust value computation in our scheme is based on two components: reputation evaluation and penalty check (see Fig.1). Reputation evaluation is based on direct and/or indirect observations, and represents the accumulative assessment of the long-term behavior, while the penalty check is based on misbehavior history that represents how much a node has misbehaved in the past.

3.2 Trust Value Computation

The calculation of a trust value needs two parts of information: direct trust value and indirect trust value. Direct trust value can be obtained when a node has direct transactions with a node. Let $T_{i,j}$ denotes the trust value from node *i* to node *j*. It is defined in (1).

$$Ti,=\alpha\ DTi,j+(1-\alpha)\ ITi,j \quad (1)$$

where $DT_{i,j}$ is the direct trust value from node i to node *j*, $IT_{i,j}$ represents the indirect trust value of node *j*, α is the confidence factor and $0 \leq \alpha \leq 1$.

A) Direct Trust Evaluation

To calculate the direct trust value, we consider two factors: the reputation rating and the penalty factor. Let $DT_{i,j}^{t}$ denotes the current direct trust value of node *j* from the view point of node *i* and $DT_{i,j}^{t-1}$denotes the past direct trust value. $Rep_{i,j}^{t}$ and $PF_{i,j}^{t}$ denote the current reputation rating and the penalty factor respectively. Therefore the trust value for node *j* at node *i* is:

$$DT_{i,j}^{t} = \begin{cases} \frac{DT_{i,j}^{t-1}+(1-PF_{i,j}^{t})}{2+PF_{i,j}^{t}} & if\ \ Rep_{i,j}^{t} = 0 \\ \frac{DT_{i,j}^{t-1}+(1-Rep_{i,j}^{t})}{2+PF_{i,j}^{t}} & otherwise \end{cases} \quad (2)$$

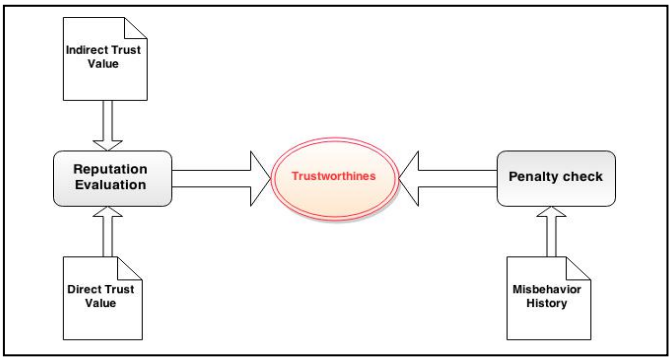


Fig. 1. Components of O^{2}Trust

If current reputation rating $Rep_{i,j}^{t}$ is equal to zero that means that the node *j* well-behaves at this moment, but there is no evidence that it is honest. To protect our trust model from on-off attacks, penalty factor that represents the misbehavior history, is used to calculate the current trust value.

In this paper, we can use one of the trust factors depending on the interactions between two neighbor nodes such as packet receive, send, delivery and consistency, to measure a node's reputation According to the quality of services provided by cooperating nodes, we classify interaction quality into two categories: successful (S) and unsuccessful (U).

In O^{2}Trust each sensor calculates individual trust values for only one-hop neighbors, contrary to GTMS [11] in which each sensor calculates individual trust values for all the cluster members. As a result, nodes do not keep trust information about every node in the network. Keeping neighborhood information implies significant lower energy consumption, less processing for trust computation, and less memory space.

Let $Rep_{i,j}^{t}$ denotes the current reputation rating which represents the current misbehavior of node *j* from the view point of node *i* at time *t*. It is defined in (3).

$$Rep_{i,j}^{t} = \frac{U_{i,j}}{U_{i,j}+S_{i,j}} \tag{3}$$

$S_{i,j}$ denotes the total number of successful interactions of node *i* with *j* during a time period *t* and $U_{i,j}$ denotes the total number of unsuccessful interactions of node *i* with *j* during a time period *t*.

Due to the uncertainty of current reputation rating value based on recent interactions experience, we introduce the penalty factor to compute the trust value and to enhance the flexibility of our trust model. Penalty factor, accumulates measured misbehavior over time. It detects the dissimulated misbehavior. So, according to our proposed method if measured misbehavior is consistent, it is always greater than predefined threshold, and each time penalty factor will be increased until it reaches to maximum value (that is one). We define the penalty factor of node j estimated by node i as follow:

$$PF_{i,j}^{t} = \begin{cases} Min\{[Rep_{i,j}^{t} + (1-\theta) \times PF_{i,j}^{t-1}], 1 \ \}, & if\ DT_{i,j}^{t} \geq THR1 \\ Min\{[Rep_{i,j}^{t} + \theta \times PF_{i,j}^{t-1}], 1 \ \}, & otherwise \end{cases} \tag{4}$$

where θ is the forgetting factor for accumulated misbehavior, which ranges from $[0.5, 1]$ and $THR1$ is a threshold that can be tuned according to the system and security requirements.

Contrary to previous trust models, in which recent rating will carry more weight and therefore past misbehavior can be completely dissimulated, in our trust model we use an adaptive forgetting factor to improve on-off attack detection. According to Equation (4) once the node's trust value is under the trust threshold $THR1$, aging factors for previous accumulative misbehavior (penalty factor) will be different. In this case, we will weigh more on the penalty factor in order to more decrease the trust value. It means the malicious node that launches on-off attack, requires a longer time to recover its trust value once it has been defined as a malicious node.

B) Indirect Trust Value

The indirect trust value is computed based on the recommendations given by neighbors when it is often not possible for a node to directly assess the trust value of another node. However, the reliability of trust and reputation models could be easily compromised by various dishonest recommendation attacks, i.e., self-promoting, bad-mouthing and collusion.

To deal with the bad-mouthing attack and collusion attack, we use a lightweight averaging function to aggregate the indirect values. So, if node *i* needs a recommendation about node *j*, it will ask only trustworthy nodes (only one-hop neighbors) in unicast mode because it is more energy efficient than broadcast mode [13]. If the direct trust of a neighbor node is larger than the trust threshold value (for example 0.6), it is declared as trustworthy neighbor.

Let us assume that be the set of the trustworthy recommenders of the node *j* defined as:

$$\psi = \{\mathrm{DT_{k,j}}, 0 \le \mathrm{k} \le \mathrm{M} - 1\ \} \tag{5}$$

were M is the total number of recommenders and $DT_{k,j}$ is the direct trust from recommender k to node j. Then the indirect trust value of node j $IT_{i,j}$ can be defined as:

$$\mathrm{IT_{i,j}} = \frac{1}{\mathrm{M}} \sum_{\substack{\mathrm{k}=0 \\ \mathrm{k}\neq\mathrm{j}}}^{\mathrm{M}-1} \mathrm{DT_{k,j}} \tag{6}$$

In [13], Liang and Shi found that the lightweight average aggregation algorithm performs better than complex algorithms.

C) Decision making

After calculating the global trust value $T_{i,j}$ that relies on [-1, 1], each node *i* will classify trust into three states as follows:

$$Mp(T_{i,j}) = \begin{cases} T\text{: trusted}, & if \quad 1 \ge T_{i,j} \ge THR1 \\ U\text{: uncertain}, & if \quad THR1 \ge T_{i,j} \ge THR2 \\ M\text{: malicious}, & if \quad THR2 \ge T_{i,j} \ge -1 \end{cases} \tag{7}$$

where $THR2 < THR1 < 1$ and $THR1$, $THR2$ are a threshold that can be tuned according to the system and security requirements to determine the node's status. Since these values depend on network and security requirements, it will be set accordingly.

According to the trust state, each node can make a decision to cooperate or non-cooperate with the interacted node in the considered operation.

4 Performance Evaluation

In this section, we present results of our simulations showing the effectiveness of our trust model. MATLAB software is used as simulation tool to assess the performance of our model. A comparative study between O^{2}Trust, *RFSN* [9] and *Retrust* [12] is given.

Concrete simulation scene is a square area of 100 m x 100 m, with 100 randomly deployed nodes. The communication radius is 25 m. An optimistic initialization strategy of trust value is adopted. So, the initial trust state of nodes is set as trusted (i.e., with initial trust value equal to 0.8).

Simulation is set up as follows. Each sensor node SN randomly selects one of its one-hop neighbors to transmit packets. Suppose that SN *i* ask SN *j* to forward packets, SN *i* can observe how many packets *j* has forwarded, i.e number of successful transactions. Next, SN *i* compute its direct trust value $DT_{i,j}^{t}$ according to equation (2). We can summarize the simulation parameters in Table 1.

Table 1. Simulation parameters

Parameter	Value	Description
$\alpha, (1-\alpha)$	(0.8, 0.2)	Weight ratio of direct and indirect value
θ	0.6	Forgetting factor
THR1	0.6 or 0.7	Trust threshold (for trusted nodes)
THR2	0,4	Trust threshold (for malicious nodes)
Initial trust value	0.8	The value assigned to a new node.

In on-off attack, strategic malicious nodes behave well and badly alternatively with the aim of remaining undetected while causing damage. Unfortunately, these malicious nodes may suddenly conduct attacks as they accumulate higher trust value. Thus, the attack cycle consists of two periods: on period and off period. When the attack is on, malicious node launches attacks; i.e. drops the received packets, and during the off period, performs well, i.e forwards received packets. Since the on period has an implication on the trust value of the malicious node, it will try to increase its trustworthiness during the off period.

4.1 Analysis of Penalty Factor Impact

In this section, we analyze the property of our trust model that combines penalty factor with reputation evaluation to derive trust value. We must demonstrate that the penalty factor helps to perceive the dissimulated misbehavior in on-off attack.

Our scheme has a feature whereby it continuously decreases the trust value of a malfunctioning or malicious node when it misbehaves in a repetitive manner. In order to validate the effectiveness of penalty factor and its influence on trust computation, we consider the actions of two types of nodes in the network: the benevolent nodes and the malicious ones. The benevolent nodes are the nodes that always behave well. While the malicious nodes are nodes that persistently misbehave.

The trust value's evolution of benevolent nodes and malicious nodes in O^2Trust is shown in fig. 2. In this experience, we calculate the average of trust values of fifty nodes of each type (benevolent and malicious). We can see that the trust value of the benevolent nodes in O^2Trust increases constantly. The factor penalty has no effect on the trust value since the behavior of trusted node is always good. However, the trust value's evolution of the malicious nodes decreases constantly as long as the malicious node persists in its misbehavior. We can see in the Fig.2, that in the first off period of attack (between 0 and 15 time units), the malicious node behaves well and its trust value follows the same evolution of the benevolent trust value. However, in the first on period (between 15 and 20 time units), it triggers the attack and its trust value falls off sharply. Consequently, its trust status changes from trusted to malicious in three time units. Since our proposed model always decreases the trust value of malicious node, the recovery rate in the off period is slower when the trust value is under the trust threshold. On the other hand, in the second on period (between 40 and 45 time units), the trust status changes from trusted to malicious in two time units. This can be explained by the fact that its last misbehavior is taken into account and as long as the malicious nodes repeat the on period, the penalty factor influences the trust value by checking the accumulated misbehavior in the past. So, it is difficult to the malicious node to recover its trust value in the off period, because the frequency of its past misbehavior is not discarded like in the previous trust models.

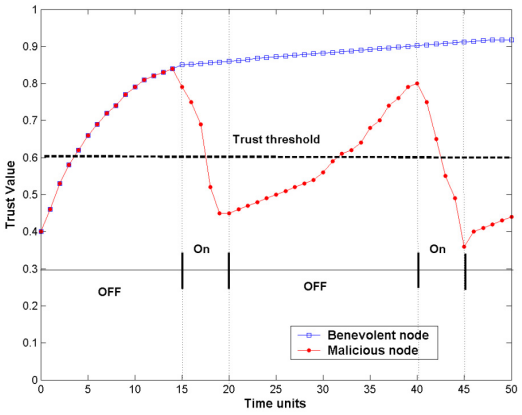


Fig. 2. Influence of penalty factor on trust computation

Consequently, considering the penalty factor in trust computation can effectively make the trust model more sensitive to on-off attack.

4.2 On-off Attack Resilience

To evaluate how our trust model can mitigate on-off attack, we introduce the malicious detection rate metric called MDN that is defined as equation (8):

$$MDN = \frac{|D|}{|M|} \tag{8}$$

Where $|D|$ denotes the number of detected malicious nodes and $|M|$ denotes the number of total malicious nodes. It is typically used to evaluate the efficiency of a trust model.

Values of the system parameters such as trust threshold and forgetting factor, are selected based on heuristic and previously defined values in the literature [11, 14, 15, 16].

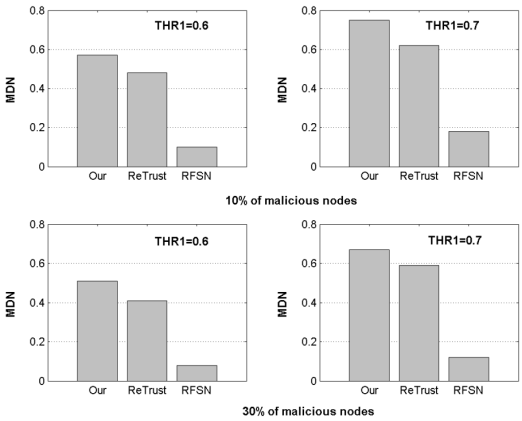


Fig. 3. Detection rate of on-off attack

Fig. 3 depicts the detection rate of on-off attack under two trust threshold values: THR1=0.6 and THR1=0.7. For each trust threshold, we consider 10% and 30% of on-off attacker nodes among 100 nodes in the network. We can clearly see that our trust model outperforms ReTrust and RFSN. While considering 10% of malicious nodes, the detection rate in O^2Trust remains 57% and 75% with the trust threshold equal to 0.6 and 0.7 respectively. However, when the proportion of malicious nodes is equal to 30%, the detection rate of O^2Trust decreases quietly and remains 51% and 67% with the trust threshold equal to 0.6 and 0.7 respectively. This is a satisfactory detection rate in trust management. On the other hand, MDN of RFSN is very lower because it cannot efficiently deal with this kind of attack and cannot recognize malicious nodes sensitively since it focus on recent behavior. Therefore, the past misbehavior is

discarded. We can also notice that when trust threshold is high, the on-off attack detection rate is also high. However, nodes might be assessed as untrustworthy even though they might not actually be malicious nodes.

We can conclude that O^2Trust is a fine-grained trust model that can portray unpredictable behaviors from malicious nodes and outperforms RFSN and ReTrust scheme. Consequently, on-off attack can be mitigated efficiently.

5 Conclusion

Trust systems are very useful mechanisms to thwart insider attacks. However, building a robust trust model is very challenging, because malicious nodes participate in the behavior rating process and can distort the trust value by cheating. In this paper, we proposed O^2Trust, a trust model to mitigate on-off attack. O^2Trust adopts the Penalty Policy against the misbehavior history of each node in the network. By considering misbehavior history, it is difficult to a malicious node to recover its trust value as long as it persists in its misbehavior. Simulation results show that O^2Trust is an efficient and on-off attack-resistant trust model. However, how to select the proper value of the weight and the defined threshold is still a challenge problem, which we plan to address in our future research endeavors.

References

1. Akyildiz, I.F., Weilian, S., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Communications Magazine 40(8), 102–114 (2002)
2. Krau, C., Schneider, M., Eckert, C.: On handling insider attacks in wireless sensor networks. Information Security Technical Report 13(3), 165–172 (2008)
3. Han, G., Jiang, J., Shu, L., Niu, J., Chao, H.C.: Management and applications of trust in Wireless Sensor Networks: A survey. Journal of Computer and System Sciences 80(3), 602–617 (2014)
4. Labraoui, N., Gueroui, M., Aliouat, M., Petit, J.: Reactive and adaptive monitoring to secure aggregation in wireless sensor networks. Telecommunication Systems 54(1), 3–17 (2013)
5. Boukerche, A., Ren, Y.: A trust-based security system for ubiquitous and pervasive computing environments. Computer Communications 31, 4343–4351 (2008)
6. Mármol, F.G., Pérez, G.M.: Providing trust in wireless sensor networks using a bio-inspired technique. Telecommunication Systems 46(2), 163–180 (2010)
7. Lopez, J., Roman, R., Agudo, I., Fernandez-Gago, C.: Trust management systems for wireless sensor networks: Best Practices. Computer Communications 33(9), 1086–1093 (2010)
8. Alzaid, H., Alfaraj, M., Ries, S., Jøsang, A., Albabtain, M., Abuhaimed, A.: Reputation-based trust systems for wireless sensor networks: A comprehensive review. In: Fernández-Gago, C., Martinelli, F., Pearson, S., Agudo, I. (eds.) Trust Management VII. IFIP AICT, vol. 401, pp. 66–82. Springer, Heidelberg (2013)
9. Ganeriwal, S., Srivastava, M.: Reputation-based framework for high integrity sensor networks. ACM Transactions on Sensor Networks (TOSN) 4(3) (2008)
10. Chen, H.: Task-based trust management for wireless sensor networks. International Journal of Security and Its Applications 3(2), 21–26 (2009)

11. Shaikh, R.A., Jameel, H., d'Auriol, B.J., Lee, H., Lee, S., Song, Y.J.: Group-based trust management scheme for clustered wireless sensor networks. IEEE Transactions on Parallel and Distributed Systems 20(11), 1698–1712 (2009)
12. Daojing, H., Chun, C., Chan, S., Bu, J., Vasilakos, A.V.: ReTrust: attack-resistant and lightweight trust management for medical sensor networks. IEEE Transactions on Information Technology in Biomedecine 16(4), 623–632 (2012)
13. Liang, Z., Shi, W.: Analysis of recommendations on trust inference in open environment. Performance Evaluation 65(2), 99–128 (2008)
14. Yu, H., Shen, Z., Miao, C., Leung, C., Niyato, D.: Survey of trust and reputation management systems in wireless communications. Proceeding of IEEE 98(10), 1755–1772 (2010)
15. Bao, F., Chen, I.R., Chang, M.J., Cho, J.: Trust-Based Intrusion Detection in Wireless Sensor Networks. In: Proceedings of IEEE International Conference on Communications (ICC), pp. 1–6 (2011)
16. Sun, Y.L., Zhu, H., Liu, K.J.R.: Defense of Trust management vulnerabilities in distributed networks. IEEE Communication Magazine 46, 112–119 (2008)

A Real-Time PE-Malware Detection System Based on CHI-Square Test and PE-File Features

Mohamed Belaoued(✉) and Smaine Mazouzi

Department of Computer Science
Université 20 août 1955-Skikda, Algeria
{m.belaoued,s.mazouzi}@univ-skikda.dz

Abstract. Constructing an efficient malware detection system requires taking into consideration two important aspects, which are the accuracy and the detection time. However, finding an appropriate balance between these two characteristics remains at this time a very challenging problem. In this paper, we present a real-time PE (Portable Executable) malware detection system, which is based on the analysis of the information stored in the PE-Optional Header fields (PEF). Our system used a combination of the Chi-square (KHI²) score and the Phi (φ) coefficient as feature selection method. We have evaluated our system using Rotation Forest classifier implemented in WEKA and we reached more than 97% of accuracy. Our system is able to categorize a file in 0.077 seconds, which makes it adequate for real-time detection of malware.

Keywords: Malware · Malware analysis · Chi-square test (KHI²) · PE-optional header

1 Introduction

Malware, abbreviation for 'malicious software', is a term used to designate any computer program that is designed to accomplish unauthorized actions without the user's consent. The number of new discovered malware has grown steadily over the past ten years. Therefore, it is crucial to have an efficient protection against this kind of malicious programs. The existing anti-malware techniques can be broadly classified in three classes, which are signature-based, behavioral-based and heuristic-based techniques [1]. Signature-based techniques are widely used by most of commercial antivirus software (AV). These techniques are very accurate for detecting known malware that exist in the signatures' database [1]. However, they are not able to deal with unknown malware or newly launched ones, often developed after discovering a zero-day exploit [2]. Even if the recent AV have become more accurate, they are still very slow to take countermeasures when a new threat is discovered [3, 4].

The behavioral analysis also known as dynamic analysis consists of monitoring the execution of the analyzed program in an isolated environment (i.e. Sandbox or virtual machine) [5]. During the monitoring process, the actions that the program accomplishes (such as API calls, Systems calls, network traffic, etc.) are recorded and used

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 416–425, 2015.
DOI: 10.1007/978-3-319-19578-0_34

to generate behavior features for categorizing the program (malware or benign). Such techniques are very accurate and they are able to detect unknown malware [5]. However, their main drawback is that the monitoring process is run for a couple of minutes at most, therefore it can't observe the entire capabilities of the program [4]. Moreover, the time required for the monitoring process makes such techniques not suitable for real-time detection.

The heuristic-based analyses investigate different file features such as Opcode instructions, structural information (Such as header information), and API (Application Programming interface) calls [1],[5]. These sets of information are used as features for the classification process, which is generally done using machine learning-based classifiers such as decision trees and Bayes Algorithm [1],[6]. The Heuristic based Anti-malware systems are very accurate and are able to deal with unknown malware [1],[5, 6]. They are also easy to implement compared to the behavioral ones. However, the existing systems suffer from the inconvenient of their high processing overhead, since most of them use a large number of features, which yields to intensive computations. Due to that, most of the existing heuristic techniques are inadequate for real-time detection, which is a very suitable characteristic especially in such sensitive systems.

In this paper, we introduce a real-time PE (Portable Executable, See section 2) malware detection system, which consists of three different components, which are the PE-parser, feature selection module and a decision module. The PE parser was developed using Python language, and it statically (i.e. without executing the analyzed program) extracts the information contained in the PE-Optional header fields (PEF, see Section 2). PE header information (including Optional header ones) are very quick to extract, which is convenient for our real-time purposes. For the same purposes, our analysis was restricted on the Optional header only. We believe that using other types of features such as File-header fields or other structural information will considerably increase the number of features, which will have a direct impact on the detection time. The feature selection module was also developed using Python, and it is based on the KHI² test, which is a statistical method used for hypothesis testing [7]. The decision module is based on Rotation Forest classifier [8] that is available in Waikato Environment for Knowledge Analysis (WEKA) [9].

This paper is organized as fellows: Section 2 introduces the PE file format in order to facilitate the comprehension of the rest of the sections. Section 3 is devoted to most known related works, published in the literature. In section 4, we present our proposed system's architecture. In section 5, we present our experimental results. Section 6 concludes our work and underlines its perspectives.

2 PE File Format

PE is an abbreviation for Portable Executable[10], and it represents the common file format for binary executables and DLLs under Windows operating systems. A PE file is structured in layers and it is mainly composed of a DOS Header, PE Header, Section Headers (Section table), and a number of sections, as shown in "figure 1".

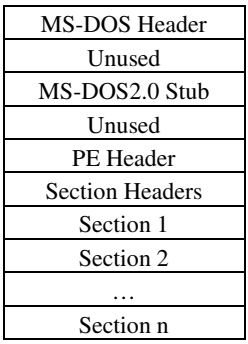


Fig. 1. PE file format

- The DOS Header is used if the file is run from the DOS. So it can then check whether it is a valid executable or not.
- The PE header is an IMAGE_NT_HEADERS data structure, which contains three members: PE-Signature, File Header, and the Optional Header. This latter is the subject of our work and is composed of several fields [10] as illustrated in figure 2. The values of the latter fields will be used as discriminators for the benign-malware categorization process.

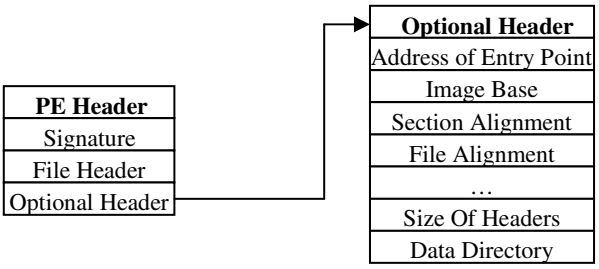


Fig. 2. Members of the PE-Optional Header

3 Related Work

In the last decade, security researchers have introduced new malware detection methods, in order to overcome the limitations of the standard signature-based ones. Schultz et al. [11] were the first authors to introduce a machine learning based malware detection system. The proposed system is based on the analysis of different information contained in the PE file such as strings and API calls. They used a classification method based on Naïve Bayes, and they achieved 97.11% of accuracy.

The method presented in [12] is based on API calls and Naïve Bayes classifier. The extracted APIs were used to construct models of suspicious behaviors, by grouping some APIs according to scenarios that a malware can accomplish, such as obtaining the system's directory, writing malicious data into files, and registry updates. They achieved an overall accuracy of 93.7%.

Ye et al. [2] have introduced a malware detection system that is based on the analysis of the set of APIs called by PE programs. The authors proposed a feature selection method based on the KHI² test. They used an Object Oriented Association (OOA) mining based classification method. Their system achieved an overall accuracy of 67.5% and a detection time of 0.09s.

The system proposed by Salehi et al. [4] is based on analyzing API calls and their arguments. They trained their system using different classifiers and they have obtained an overall accuracy of 98.1%. Extracting APIs arguments requires executing the program; therefore, this method has the inconvenient of dynamic approaches mentioned previously.

4 Proposed Method

Our proposed malware detection system categorizes a file in three different phases, which are the feature extraction, the feature selection, and the decision (classification).

4.1 Feature Extraction

As mentioned previously, our system relies on the analysis of the PE Optional Header fields (PEFs) and in order to extract these features from the analyzed file we developed a module written in Python by using a third party Python module called pefile [13]. PEFs are generated by concatenating the field's name and value (ex. CheckSum0 designates that the feature CheckSum has a value equal to 0).

4.2 Feature Selection

In order to reduce the number of obtained PEFs and keep only the most relevant ones, we developed a feature selection method, which is based on the chi-square (KHI²) test. The KHI² is a statistical method, which is used to determine whether there is a significant association between two qualitative variables. This association is expressed by the distance $\boldsymbol{D}$ between an observed frequency $\boldsymbol{O}$ and an expected one $\boldsymbol{E}$ (which represents the case of independence between the variables) and the greater is that distance stronger is the correlation between the variables. In our case, we will study that association between the variable 'PEF' that has two modalities: "present" and "absent". This variable represents the presence or not of a specific PEF in a PE file. The second variable is "PE" that has also two modalities: "Malware" and "Benign" that corresponds to the two categories of PE files that we used.

The first step to do when conducting a KHI² test, is to define the two hypotheses $\mathbf{H_0}$ and $\mathbf{H_1}$ that one will be accepted, and the other rejected. $\mathbf{H_0}$ and $\mathbf{H_1}$ represent respectively the case of independency and the case of dependency between the two variables. Note that accepting $\mathbf{H_0}$ for a PEF means that it is not specific to any category of PE-files. Therefore, it will be considered as irrelevant and will be removed. In our case, $\mathbf{H_0}$ and $\mathbf{H_1}$ are defined as follows:

- **H_0**: The presence or absence of a PEF is independent of the PE file's type (malware or benign).
- **H_1**: The presence or absence of PEF is related to the PE file's type (malware or benign).

For every PEF, we have a contingency table as shown in table 1.

Table 1. Contingency Table of a PEF.

	PEF: Present	**PEF: Absent**	**Row Total**
PE: Malware	N1	N2	N
PE: Benign	M1	M2	M
Column total	N1+M1	N2+M2	T

N, M, and ***T*** are respectively the total number of malware PE, the total number of benign PE, and the total number of all PE files ($T=N+M$). ***N1*** and ***N2*** are respectively the number of malware PE that have a **PEF** and the number of malware PE that do not have the **PEF**, such as ***N = N1 +N2***. ***M1*** and ***M2*** are respectively the number of benign PE that have a **PEF** and the number of benign PE that do not have the **PEF**, such as ***M = M1 + M2***. The KHI² score (D^2) is calculated using the formula (1):

$$D^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \tag{1}$$

Where $O_{r,c}$ is the observed frequency count at level r of row variable and level c of column variable. And $E_{r,c}$ is the expected frequency. $E_{r,c}$ is defined by equation (2) .

$$E_{r,c} = \frac{n_r \times n_c}{T} \tag{2}$$

Where n_r, and n_c represent respectively the sum on row r and the sum on column c. After calculating the KHI² values for the obtained PEFs, we have to determine which of two hypotheses are accepted or rejected for every PEF. To do that, we have to compare the obtained KHI² scores of every PEF to a threshold, which represents the theoretical KHI² value (χ^2). That value is obtained by first calculating the degree of freedom (**DF**), and choosing a signification level α that represents the error probability when accepting or rejecting an hypothesis. Considering **DF** and α, the χ^2value is obtained from the KHI² distribution table [14] . **DF** is calculated using the following equation:

$$DF = (R-1) \times (C-1) \tag{3}$$

Where R, and C are respectively the number of modalities of the first and the second variables. After rejecting all the PEFs that are not correlated ($D^2 \leq \chi^2$), we will calculate the φ coefficient using the formula (4) for the remaining ones. The φ coefficient is a normalization of the KHI² score (D^2), which is used to measure the strength of the dependency between the two variables [15]. In our work, that coefficient will be used to generate the different PEFs' subsets, which are grouped according to their correlation's strength (relevance).

$$\varphi = \sqrt{\frac{D^2}{T}} \tag{4}$$

The value of φ ranges between 0 and 1, therefore, the strength of the relationship can be divided in 4 different classes:

- $\varphi \approx 0.25$: Weak correlation.
- $\varphi \approx 0.50$: Medium correlation.
- $\varphi \approx 0.75$: Strong correlation.
- $\varphi \approx 1$: Very strong correlation.

Our obtained PEFs will be divided into non-disjoints subsets according to the φ values mentioned previously.

4.3 Classification

In order to evaluate our malware detection system we have used Rotation Forest classifier [8] that is implemented in WEKA [9]. Therefore, our classification module takes as an input the PEFs subsets represented as an .arff file. The .arff file is the data file format supported by WEKA, and it is automatically generated using a python script. The classifier is then trained and models are generated for each feature subset of the training set. The obtained models are then tested on previously unseen PE-files contained in our test set.

5 Experimentation

5.1 Dataset

We collected a dataset composed of 552 PE files (338 malware and 214 benign programs). This dataset will be split into 80% training set and 20% test set. The infected PE dataset was downloaded from Vxheavens.com and contains 12 different malware categories as shown in table 2.

Table 2. Used malware dataset

N°	Malware Type	Counts	N°	Malware Type	Counts
1	Backdoor	27	7	Trojan	59
2	Email-Worm	19	8	Trojan-Downloader	24
3	Exploit	28	9	Trojan-Dropper	32
4	Hacktool	22	10	Trojan-Spy	18
5	Net-Worm	16	11	Virus	42
6	P2P-Worm	17	12	Worm	34
TOTAL = 338					

The benign PE files include some utility software downloaded from Softpedia.com and some Windows system files collected from a clean installation of windows XP. We scanned the whole dataset by more than 40 AV available on the website Virus-Total.com, in order to make sure that they are correctly labeled (malware, benign).

5.2 Results and Evaluation

In this subsection, we will present the obtained experimental results from the feature extraction phase until the decision phase. We first start by the obtained PEFs after the feature extraction phase. As presented in Table 3, we have obtained 590 PEFs with their corresponding frequencies in malware and benign PE (observed frequencies).

Table 3. Overview of the obtained PEFs list and their corresponding frequencies

			Frequency	
N°	**Optional Header field**	**Value**	**Malware (271)**	**Benign(172)**
1	BaseOfCode	4096	271 (100%)	172 (100%)
2	BaseOfData	102400	4 (1%)	1 (1%)
…	…	…	…	…
86	CheckSum	0	259 (96%)	5 (3%)
87	CheckSum	102910	1 (1%)	0 (0%)
…	…	…	…	…
589	Subsystem	2	226 (83%)	106 (62%)
590	Subsystem	3	45 (17%)	66 (38%)

We will calculate the KHI^2 and φ values (as presented in the subsection 4.2) for the obtained PEFs and remove the non-relevant ones that have $KHI^2 < 3.84$ (3.84 is the X^2 value for DF=1 and alpha =0.05). The obtained results are presented in table 4.

Table 4. KHI^2 scores and φ values of the selected PEFs

N°	**PEF**	***KHI²***	***φ***
1	CheckSum0	375.21	0.92
2	MajorImageVersion0	370.57	0.91
3	DllCharacteristics0	355.91	0.9
4	MajorOperatingSystemVersion5	346.02	0.88
5	MinorOperatingSystemVersion0	341.92	0.88
…	…	…	…
50	SizeOfInitializedData28672	3.86	0.09

As presented in Table 4, we have obtained a final list of 50 PEFs with their corresponding KHI^2 scores and φ values. We will divide these features into different groups (subsets) according to their φ values. At the end of the feature selection phase, we have obtained three different subsets: G1, G2, and G3 that contain PEFs that have

respectively $\varphi \geq 0.75$, $\varphi \geq 0.5$, and $\varphi \geq 0.25$. We have respectively 11, 14, and 22 PEFs in G1, G2, and G3. We have used a forth subset G4 that contains the complete 590 extracted PEFs, the aim from that is to see whether our feature selection method have improved the obtained results or not.

Next, we will evaluate our system's performance by training the Rotation Forest classifier using different features subsets and see which subset will generate the best results. The performance of a classifier is generally evaluated by calculating three different metrics which are Detection rate (DR), False Alarm rate (FA), and Accuracy (AC) and they are calculated using the equations 5, 6 and 7 respectively:

$$DR = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

Where TP (true positive) and FN (false negative) represent respectively malware that were correctly classified as malware and malware that were wrongly classified as benign.

$$FA = \frac{FP}{FP + TN} \times 100\% \tag{6}$$

TN (true negative) and FP (false positive) represent respectively benign programs that were correctly classified as benign, benign program that were wrongly classified as malware. The accuracy (AC) represents the rate of files that were correctly classified in their class.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{7}$$

The fourth metric that we will use to evaluate our system's performance is the detection time (DT), which represents the average time required for categorizing a file and it is expressed in seconds per file. DT includes the feature extraction time, .arff file generation time, and the classification time. The obtained results are presented in table 5.

Table 5. Experimental results

Group	φ	PEF Counts	DR	*FA*	AC	DT
G1	≥0.75	11	98.51%	7.14%	96.33%	0.075
G2	**≥0.50**	**14**	**100.00%**	**7.14%**	**97.25%**	**0.077**
G3	≥0.25	22	98.51%	9.52%	95.41%	0.079
G4	-	590	97.01%	7.14%	95.41%	0.116

From the results presented in the above table, we can see that our proposed feature selection method was able to increase the accuracy of our system by +1.84% (from 95.41% with G4 to 97.25% with G2) and that using only 14 PEFs. It was also able to reduce the categorization time by 33% (from 0.116s with G4 to 0.077s with G2). Note that the feature extraction phase took 0.037s, the .arff file generation also required 0.037s, and the classification phase took 0.003s.

5.3 Comparison

In this subsection, we will evaluate our system's performance by comparing it with the previously cited methods. The results are presented in table 6.

Table 6. Results of the comparison with the previously cited methods for malware detection

Method	Feature Type	DR	AC
Our method	**PEFs**	**100%**	**97.25%**
Schultz et al. [11]	Strings	97.43%	97.11%
Salehi et al. [4]	APIs+Args	99.2%	98.4%
Wang et al. [12]	APIs	94.4%	93.71%
Ye et al. [2]	APIs	88.16%	67.5%

From the results presented in Table 6, we can see that our system outperforms three of the four presented systems with an improvement in accuracy that varies from 0.14 % to 30%. The system proposed by Salehi et al. [4] is more accurate than our system (+1.15%). However, our system has a better detection rate.

If we consider the detection time (categorization time), we can conclude that our system is adequate for real-time detection. The proposed system is able to categorize a file in 0.077s, which is a very satisfying performance, compared with the system proposed by Ye et al. [2] which categorizes a file in 0.09s. The system proposed by Salehi et al. [4] needs to monitor the analyzed program during 2 minutes in order to extract API calls and their arguments, that represents almost 3000 times the required time by our proposed features extraction method.

6 Conclusion and Future Works

In this paper, we have presented a real-time PE-malware detection system that is based on the analysis of the PE-optional Header information. The proposed system uses an efficient feature selection method, which is based on the KHI² test. This latter allowed us to achieve a high accuracy and a low detection time, using only 2% of the initially extracted features. As future works, we project to combine different types of features such as APIs calls, and Opcode, in order to increase the accuracy of our system.

References

1. Bazrafshan, Z., Hashemi, H., Fard, S.M.H., Hamzeh, A.: A survey on heuristic malware detection techniques. In: Proceedings 2013 5th Conference on Information and Knowledge Technology (IKT), Shiraz, pp. 113–120 (2013)
2. Ye, Y., Li, T., Jiang, Q., Wang, Y.: CIMDS: Adapting postprocessing techniques of associative classification for malware detection. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 40, 298–307 (2010)
3. June, I.: Anti-malware vendors slow to respond. Computer Fraud & Security, 1–2 (2010)

4. Salehi, Z., Sami, A., Ghiasi, M.: Using feature generation from API calls for malware detection. Computer Fraud & Security Bulletin, 9–18 (2014)
5. Aycock, J.D.: Computer viruses and malware. Springer, Heidelberg (2006)
6. Shabtai, A., Moskovitch, R., Elovici, Y., Glezer, C.: Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey. Information Security Technical Report 14, 16–29 (2009)
7. Fornasini, P.: The Chi Square test. The Uncertainty in Physical Measurements:An Introduction to Data Analysis in the Physics Laboratory, pp. 187–198. Springer Science & Business Media (2009)
8. Rodríguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A New classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1619–1630 (2006)
9. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
10. Pietrek, M.: Peering Inside the PE: A Tour of the Win32 Portable Executable File Format. Microsoft Systems Journal-US Edition 9, 15–38 (1994)
11. Schultz, M.G., Eskin, E., Zadok, E., Stolfo, S.J.: Data mining methods for detection of new malicious executables. Proceedings. In: 2001 IEEE Symposium on Security and Privacy, S&P 2001, Oakland, CA, pp. 38–49 (2001)
12. Wang, C., Pang, J., Zhao, R., Liu, X.: Using API sequence and bayes algorithm to detect suspicious behavior. In: Proceedings of the 2009 International Conference on Communication Software and Networks, ICCSN 2009, Macau, pp. 544–548 (2009)
13. `https://code.google.com/p/pefile/`
14. Koskska, S., Nevison, C.: Statistical tables and formulae. Springer, New York (1989)
15. Farrington, D.P., Loeber, R.: Relative improvement over chance (RIOC) and phi as measures of predictive efficiency and strength of association in 2x2 tables. Journal of Quantitative Criminology 5, 201–213 (1989)

Security and Network Technologies: Wireless Sensor Networks

Balanced and Safe Weighted Clustering Algorithm for Mobile Wireless Sensor Networks

Amine Dahane(✉),Nasr-Eddine Berrached, and Abdelhamid Loukil

Intelligent Systems Research Laboratory (LARESI)
University of Sciences and Technology of Oran
P.O. Box 1505, Oran, Algeria
amineusto.laresi@gmail.com

Abstract. The main concern of clustering approaches for mobile wireless sensor networks (WSNs) is to prolong the battery life of the individual sensors and the network lifetime. In this paper, we propose a balanced and safe weighted clustering algorithm which is an extended version of our previous algorithm (ES-WCA) for mobile WSNs using a combination of five metrics. Among these metrics lie the behavioral level metric which promotes a safe choice of a cluster head in the sense where this last one will never be a malicious node. The goals of the proposed algorithm are: offer better performance in terms of the number of re-affiliations which enables to generate a reduced number of balanced and homogeneous clusters, this algorithm, coupled with suitable routing protocols, aims to maintain stable clustering structure. We implemented and tested a simulation of the proposed algorithm to demonstrate its performance.

Keywords: WSNs · Clustering · Homogenous Clusters · Energy Efficiency · Security

1 Introduction

After the success of theoretical research contributions in previous decade, wireless sensor networks (WSNs) [1,2] have become now a reality. Their deployment in many societal, environmental and industrial applications makes them very useful in practice. These networks consist of a large number of small size nodes which sense ubiquitously some physical phenomenon (temperature, humidity, acceleration, noise, light intensity, wind speed, etc.) and report the collected data to the sink station by using multi-hop wireless communications. The clustering concept, that means grouping nodes which are close to each other, has been studied largely in ad-hoc networks [2,3,4,5,6,7,8] and recently in WSNs [9,10,11,12,13] where the purpose in general is to reduce useful energy consumption and routing overhead, however, cluster-heads must be selected carefully and diligently. Recent research studies recognize that organizing mobile WSNs, in the sense defined above, into clusters by using a clustering mechanism is a challenging task [9,10]. This is due to the fact that cluster heads (CHs) carry out extra work, and consequently consume more energy compared with cluster members (CMs) during the network operations and this will lead to untimely death causing network partition and therefore failure in communication link. For this reason, one of the frequently encountered problems in this mechanism is to search for the best way to elect CH for each

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 429–441, 2015.
DOI: 10.1007/978-3-319-19578-0_35

cluster. Indeed, a CH can be selected by computing quality of nodes, which may depend on several metrics: connectivity degree, mobility, residual energy and distance of a node from its neighbors. Significant improvement in performance of this quality can be achieved by combining these metrics [2,3,8,9,13,14].

In this paper, we propose balanced and safe weighed clustering algorithm for mobile WSNs (BS-WCA) using a combination of the above metrics with the behavioral level metric which we have added. Our approach enables to generate a reduced number of balanced and homogeneous clusters in order to minimize the energy consumption of the entire network and prolong sensors lifetime. In the other sense, the behavioral level is decisive and allows the proposed clustering algorithm to avoid any malicious node in the neighborhood to become a CH, even if the remaining metrics are in its favor. The election of CHs is carrying out using weights of neighboring nodes which are computed based on selected metrics. So, this strategy ensures the election of legitimate and trustworthy CHs with high weights. The Node-Weight heuristic assigns node-weights based on the suitability of nodes acting as cluster heads and the election of the cluster head is done on the basis of the largest weight among its neighbors. This means that a node decides to become a cluster head or stay as an ordinary node depending on the weights of its one hop neighbors [2].

The preliminary results obtained through simulation study demonstrate the effectiveness of our algorithm in terms of number of equilibrate clusters, number of re-affiliations, by comparing it with WCA [2], DWCA [14] and SDCA [11].

These results also reveal that our approach is very suitable if we plan to use in network layer reactive routing protocols instead of proactive ones after the clustering mechanism was launched. The contribution of our paper is as follows:

- Maintaining stable clustering structure and offering better performance in terms of the number of re-affiliations using the proposed algorithm BS-WCA.

The remaining part of this paper is organized as follows: We first, in Section 2, discuss the existing studies. The details of our approach are described in section 3. Section 4 introduces and explains the selected metrics for the proposed approach of clustering. A special attention was reserved for this last aspect in this research. More details on the proposed algorithm are provided in section 5. Section 6 presents the simulation tool developed for the evaluation and provides simulation results to show the effectiveness of the proposed algorithm. Section 7 concludes the paper.

2 Related Works

In this section, we outline some approaches of clustering used in Ad-hoc networks and WSNs. Abbasi *et al.* [15] presented taxonomy and classification of typical clustering schemes, and then summarized different clustering algorithms for WSNs based on classification of variable convergence and constant convergence time protocols. They also highlighted objectives, features, and algorithms complexity. Research studies on clustering in Ad-hoc networks evolve surveyed works on clustering algorithms [16] and cluster head election algorithms [3,10]. For the single metric based on clustering, as in paper [17], the node with the least stability value is elected as CH among its neighbors, however the choice of CH which has a lower energy level, could quickly become a bottleneck of its cluster. Safa *et al.* [4] designed and implemented a dynamic energy efficient clustering algorithm (DEECA) for mobile Ad-hoc networks (MANETs) that increases

the network lifetime, however, the cluster formation in this scheme is not based on connectivity so the formed clusters are not well connected; this induces an increase of re-affiliation rate and re-clustering situations. Other proposals use strategy based on weights computing in order to elect CHs [2,3,8,14].The main strategy of these algorithms is based mainly on adding more metrics such as connectivity degree, mobility, residual energy and distance of a node from its neighbors, corresponding to some performance in the process of electing CHs. Although, the algorithms using this strategy allow to ensure the election of a better CHs based only on their high weights computed from the considered metrics, but unfortunately they does not ensure that the elected CHs are legitimated nodes, which is to say if the election process of CHs is safe or not. Safa *et al.* [5] propose a novel cluster-based trust-aware routing protocol (CBTRP) for MANETs to protect forwarded packets from intermediary malicious nodes. The proposed protocol ensures the passage of packets through trusted routes only by making nodes monitor the behavior of each other and update their trust tables accordingly. However, in CBTRP all nodes monitor the network which lead rapid drainage of node energy and therefore minimize the lifetime of the network. Khalil *et al.* [18] proposed a protocol called DICAS, which uses local monitoring and mitigates the attacks against control traffic by detecting, diagnosing and isolating the malicious nodes. Hsin *et al.* [19] proposed a self-monitoring mechanism that pays more attention to the system-level fault diagnosis of the network, especially for detecting node failures. However, they did not deal with malicious behaviors. Little effort has been made in introducing security aspect in clustering mechanism. Yu *et al.*[7] tried to secure clustering mechanism against wormhole attack in ad-hoc networks (communication between CHs) but after forming clusters, not during the election procedure of CHs. Hai *et al.*[21] propose a lightweight intrusion detection framework integrated for clustered sensor networks by using an over-hearing mechanism to reduce the sending alert packets. Elhdhili *et al.* [6] propose a reputation based clustering algorithm (RECA) that aims to elect trustworthy, stable and high energy cluster heads but during the election procedure, not after forming clusters. Benahmed *et al.* [11] used clustering mechanism based on weighted computing as an efficient solution to detect misbehavior nodes during distributed monitoring process in WSNs. However, they focused only on the misbehavior of malicious nodes and not on the nature of attacks, the formed clusters are not homogeneous, the proposed secured distributed clustering algorithm (SDCA) is not coupled with routing protocols and doesn't give much importance to energy consumption.

In the context of these surveyed research works about clustering in both ad-hoc networks and WSNs, we classified our contribution among approaches based on the computing of the weight of each node in the network, this approach focuses around strategy of distributed resolution which enables to generate a reduced number of balanced and homogeneous clusters in order to minimize the energy consumption of the entire network and prolong sensors lifetime. Moreover, we introduced a new metric (the behavioral level metric) which promotes a safe choice of a cluster head in the sense where this last one will never be a malicious node.

3 Our Approach

In the literature, no research has thought to use energy efficiency and monitoring mechanism using the same cluster-based architecture. Our first objective is to make the network able to self-organize in order to achieve its tasks with a least cost. In this context, we must determine the parameters for generating a reduced number of stable and balanced clusters. Our second objective is to propose a mechanism that assures the distributed monitoring of WSNs security reasons. This mechanism uses a cluster-based architecture, as well as new set of metrics and rules for diagnosing the state of the sensors. The advantages of this solution are that it reduces the flow of communication and provides stable surveillance environment. This approach gives more importance to the election criteria of nodes responsible for monitoring the network. The details of this approach are illustrated in our proposed algorithm BS-WCA.

4 .Metrics for CHs Election

This section introduces the different metrics used for cluster-head election. In our earlier work [9], we insisted in Mobility (Mi), connectivity (Ci), residual energy (Eri) and distance of node ni (Di) to its neighbors. In this paper, we focus our study on behavior level metric.

- **The Behavior Level of a Node n_i (BLi)**

The behavioral level of a node n_i is a key metric in our contribution. Initially, each node is assigned an equal static behavior level "BL_i=1".

However, this level can be decreased by the anomaly detection algorithm if a node is misbehavior as illustrated by Fig 1.

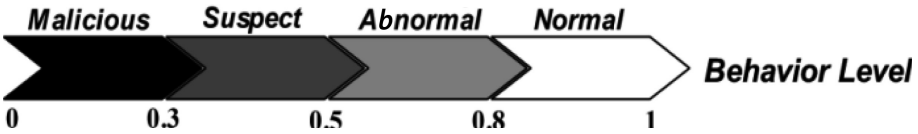


Fig. 1. The behavior level (BL_i)

For computing the behavior level of each node, nodes with a behavior level less than threshold behavior will not be accepted as CH candidates even if they have the other interesting characteristics such as high energy, high degree of connectivity or low mobility. Nevertheless, abnormal node and suspect node can always belong to a cluster as a CM but never as a CH. So, we define the behavior level of each sensor node n_i, noted BL_i, in any neighborhood of the network as presented in Fig.1. BL_i is classified by the following mapping function ($Mp(BL_i)$):

$$Mp(BL_i) = \begin{Bmatrix} Normal\ node: & 0.8 \le BL_i \le 1 \\ Abnormal\ node: & 0.5 \le BL_i < 0.8 \\ Suspect\ node: & 0.3 \le BL_i < 0.5 \\ Malicious\ node: & 0 \le BL_i < 0.3 \end{Bmatrix} \quad (1)$$

The values in the formula (1) are chosen on the basis of several reputed models of WSNs adopted by numerous researchers like Shaikh *et al.* [20] and Hai *et al.* [21].
For each node, we must calculate its weight P_i, according to the equation:

$$P_i = w_1 * BL_i + w_2 * Er_i + w_3 * M_i + w_4 * C_i + w_5 * D_i \quad (2)$$

Where $w_1, w_2, w_3, w_4,$ and w_5 are the coefficients corresponding to the system criteria, so that:

$$w_1 + w_2 + w_3 + w_4 + w_5 = 1 \quad (3)$$

We propose to generate homogeneous clusters whose size lies between two thresholds:$Thresh_{Upper}$ and $Thresh_{Lower}$. These thresholds are arbitrarily selected or depend on the topology of the network. Thus, if their values depend on the topology of the network, they are calculated as follows according to [12]:

$$Thresh_{Upper} = \frac{1}{2}(\delta_{12}(u) + AVG) \quad (4)$$

$$Thresh_{Lower} = \frac{1}{2}(\delta_{12}(v) + AVG) \quad (5)$$

With:

$$\delta_{12}(u) = \max(\delta_{12}(u_i):\ u_i \in U) \quad (6)$$

$$\delta_{12}(v) = \min(\delta_{12}(v_i):\ v_i \in U) \quad (7)$$

$$AVG = \frac{\sum_{i=1}^{n} \delta_{12}(u_i)}{N} \quad (8)$$

Where:
- u represents the node that has the maximum number of neighbors with one jump;
- v represents the node that has the minimum number of neighbors with one jump;
- AVG denotes the average cardinal of the groups with one jump of all the nodes of the network;
- N is the number of nodes in the network.

The weight P_i calculated for each sensor is based on the above parameters ($BL_i, M_i, D_i,$ Er_i and C_i). It means for our case the trust level of each node in the network. The values of coefficients w_i should be chosen depending on the importance of each metric in considered WSNs applications. For instance, we can assign a greater value to the metric BL_i compared to other metrics if we promote the safety aspect in the clustering mechanism. We can also assign a same value for each coefficient $w_i$in case when all metrics are considered having the same importance. An approach based on these weights will enable us to build a self-organizing algorithm able to form small number of homogenous clusters in size and radius by grouping geographically close nodes. The resulting weighted clustering algorithm reduces energy consumption and guaranty the choice of legitimate CHs.

5 Weighted Clustering Algorithm (BS-WCA)

In this section, we first give assumptions of the proposed algorithm: Balanced and Safe Weighted Clustering Algorithm (BS-WCA).

Then we present, in detail, an extended version of ES-WCA [9] followed by an illustrative example.

5.1 Assumptions

Before heading into the technical details of our algorithm, this paper is based on the same assumptions as in [9]. We add the fact that a malicious node can use its own ability to move freely in the space area. The behavior of the malicious node by moving frequently inside a same cluster or from a cluster to another is a normal behavior to not attract attention of the neighborhood and therefore to be detected .

5.2 Re-affiliation Phase

During the first phase, it may not be possible for all clusters to reach the $Thresh_{Upper}$ threshold. Moreover, it is possible that clusters whose size is lower than $Thresh_{Lower}$ may be created, since there is no constraint relating to the generation of these types of clusters. BS-WCA uses four types of messages in the Re-affiliation phase. The message RE_AFF_CH, that is sent in the network by the CH which the cluster size is less than $Thresh_{Upper}$. The second one is the REQ_RE_AFF message that is sent by the neighbors of CH if it wants to join this cluster. Finally a CH must send a response ACCEPT_RE_AFF message or DROP_AFF message as illustrated by Fig. 2. Hence, in this second phase, we tried to reduce the number of clusters formed and reorganize them in order to obtain balanced and homogeneous clusters. For that, we propose to re-affiliate the sensor nodes belonging to clusters that have not attained the cluster size$Thresh_{Lower}$to those that did not reach $Thresh_{Upper}$.

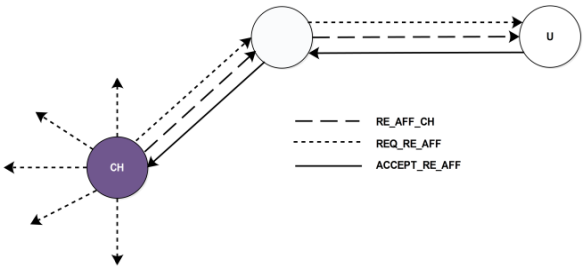


Fig. 2. Procedure of Re-affiliation of node ‘U’ to a cluster

We demonstrate our set up phase algorithm and re-affiliation phase with the help of four figures (Fig. 3, Fig. 4, Fig. 5 and Fig. 6).

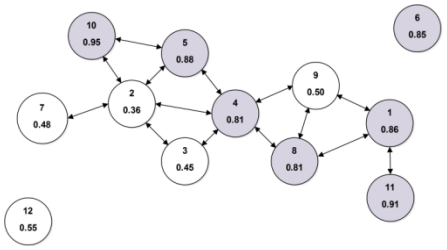


Fig. 3 Topology of the network

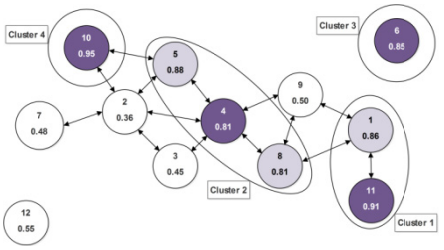


Fig. 4. Identification of clusters nodes

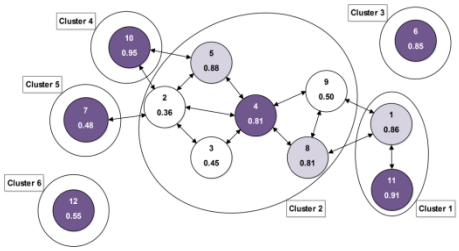


Fig. 5. The final identification of clusters

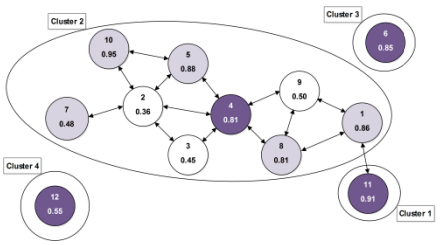


Fig. 6. The final identification of clusters

Algorithm: Re-affiliation Phase Algorithm

Inputs: $Thresh_{Upper}$, $Thresh_{Lower}$;
Outputs: set of clusters
Begin
1: For num_cl = 1 to Count (Cluster) **Do**
2: **If** (Size (Cluster [num_cl]) < $Thresh_{Upper}$)
Then
3: **C**H sends a message "RE_AFF_CH" to its neighbours (N(CH));
4: J = Count (N(CH));
EndIf
5: **For** I = 1 to J **Do**
6: **If** ($n_i \in N$(CH) receives the message)
&& ($n_i \in$ (Size(Cluster[num_cl]) <$Thresh_{Lower}$)
Then
7: n_i sends a Select message "REQ_RE_AFF" to the CH;
8: **If** (Size (Cluster [num_cl]) < $Thresh_{Upper}$)
Then
9: CH sends a message "ACCEPT_RE_AFF" to n_i;
10: CH updates its state vector;
11: CH → CH → Size = Size + 1;
12: n_i updates its state vector;
13: n_i → CH → ID = ID;
14: **Else** CH sends a "FIN_ AFF" message to n_i;
15: Go to 2;
EndIF
16: **Else** n_i sends a "DROP_AFF" message to CH;
EndIf
End For
End For
End.

Table I shows the values of the different criteria for the nodes that have behavior level $BL_i > 0.8$ (Normal nodes). Table II shows the weights P_i of neighbors for each node that have behavior level $BL_i > 0.8$.

Table 1. Values of the various criteria of normal nodes

Ids	BL_i	Eri	C_i	D_i	M_i	P_i
1	0.86	3842.12	3	1.15	1.20	769.632
4	0.81	4832.54	5	2.30	0.30	968.133
5	0.88	4053.25	3	1.30	0.55	811.829
6	0.85	4620.43	0	0.00	0.20	924.361
8	0.81	4816.80	4	1.05	1.40	964.753
10	0.95	3650.25	2	0.55	0.10	730.805
11	0.91	4819.60	1	0.70	2.20	964.753

Table 2. Weight P_i of neighbors

Ids	1	4	5	6	8	10	11
1	769.632	-	-	-	964.753	-	964.753
4	-	968.133	811.829	-	964.753	-	-
5	-	968.133	811.829	-	-	730.805	-
6	-		-	924.361	-	-	-
8	769.632		-	-	964.753	-	-
10	-	968.133	811.829	-	-	730.805	-
11	769.632		-	-	-	-	964.753

Nodes in Fig.3 are presented by circles containing their identity Ids at the top and the levels of behavior at the bottom. According to table 2, node 1 has a choice between CH11 and CH8 (they have the same weight), but the behavior level of node 11 is greater than the node 8 (BL_{11}>BL_8), so node 1 will be attached to CH11. For the other nodes, we have various conditions. Node 4 declares itself as a CH. Node 5 will be attached to CH4. Node 6 declares itself as a CH, because it is an isolated node. Node 8 will be attached to CH4. Node 10 is connected with CH5, but node 5 is attached to CH4; thus, node 10 declares itself as a CH. Node 11 declares itself as a CH. These results give us the representation shown in Fig.4. Node 2 is connected with CH4 and CH10. Node 2 will be attached to CH4, because CH4 has the maximum weight (968.133). Node 3 is connected with CH4, which implies that node 3 will be attached to CH4. Node 7 is not connected with any CH, so node 7 declares itself as CH. Node 9 is connected with CH4, and then node 9 will be attached to CH4. Node 12 is not connected with any CH, which implies that node 12 declares itself as a CH. These results give us the representation shown in Fig.5. We propose to generate homogeneous clusters whose size lies between two thresholds: $Thresh_{Upper}$= 9 and $Thresh_{Lower}$= 6. For that, we suggest to re-affiliate the sensor nodes belonging to the clusters that have not attained the cluster size $Thresh_{Lower}$ to those that did not reach $Thresh_{Upper}$. Node 4 have the highest weight and his size is less than $Thresh_{Upper}$. Nodes 1, 7 and 10 are neighbors of the node 4 with 2 hops and belong to the clusters that have not attained the cluster size $Thresh_{Lower}$, so these

nodes get merged to cluster 2. Clusters 1, 3, and 4 will be homogeneous with cluster 1 when the network becomes densely. At the end of this example, we obtain a network of four clusters (as shown in Fig. 6).

There are five situations that require the maintenance of clusters:

- Battery depletion of a node.
- Behavior level of a node less than or equal 0.3.
- Adding, moving or deleting a node.

In all of these cases, if a node n_i is CH then the set-up phase will be repeated.

6 Implementation Results

In this section, we present our simulator 'Mercury' and the results of our work. To determine and evaluate the results of the execution of algorithms that are introduced previously, the number of sensors (N) to deploy must be less than or equal to 1000. There are two types of sensor node deployment on the sensoring field: random and manual. "Mercury" offers users the ability to select a sensor type from 5 predefined types. Each one has its characteristics (radius, energy, etc.). The user can also introduce his own characteristics. The unity of the used energy is the nano joules ($1\ \text{Joule} = 10^9$ NJ).

6.1 Discussion and Results

In all experiments, N varies between 10 and 100 sensor nodes, the transmission range (R) varies between 10 and 70 meters (m) and the used energy (E) equal to 50000 NJ. By default, for each set of simulation, we conduct 100 runs with different node generations and report the average. The sensor nodes are randomly distributed in a "570m × 555m" space area by the following function:

```
for (int n = 0; n < node_tobe_deployed; n++)
{
    X_ = rand()  %  image_Field_Of_Collecting -> width;
    Y_ = rand()  %  image_Field_Of_Collecting -> Height;
}
```

To measure the performance of BS-WCA algorithm, we considered the following four metrics:

a. The number of clusters;
b. The number of re-affiliations;

The values of weighting factors used for simulation were:

$$w_1 = 0.3,\ w_2 = 0.2,\ w_3 = 0.2,\ w_4 = 0.2 \text{ and } w_5 = 0.1.$$

Note that these values are arbitrary at this time and should be adjusted according to the system requirements. To evaluate the performance of the BS-WCA algorithm with other algorithms, we studied the effect of the density of the networks (number of sensor nodes in a given area) and the transmission range on the average number of

formed clusters. Then we compare it with a DWCA (Distributed Weighted Clustering Algorithm) proposed in [14], WCA (A weighted Clustering Algorithm for Mobile Ad-hoc Networks) proposed in [2] and SDCA (secured distributed clustering algorithm) proposed in [11]. We omit presenting all results and the monitoring phase due to the space limitation. The highlight of our work is summarized in a comprehensive strategy for monitoring the network that will be presented in our future works. The goal is to detect and remove the malicious nodes

Fig.7 depicts the average number of clusters that are formed with respect to the total number of nodes in the network. The communication range used in this experience is 200m. As we can see in Fig. 7, the proposed algorithm produced the same number of clusters than DWCA when the node number is equal to 20 nodes. If the node density has increased, BS-WCA would have produced constantly less clusters than SDCA and DWCA regardless of node number. The result of BS-WCA is so unstable between 60 and 90 because we use a random deployment so if the distance between the nodes increases, the number of clusters increases too. When there were 100 nodes in the network, the proposed algorithm produced about 61.91% fewer clusters than DWCA [14] and about 38.46% than SDCA [11]. As a result, our algorithm gave better performance in terms of the number of clusters when the node density in the network is high, because BS-WCA generates a reduced number of balanced and homogeneous clusters, whose size lies between two thresholds: $Thresh_{Upper}$ and $Thresh_{Lower}$ (Re-affiliation Phase) in order to minimize the energy consumption of the entire network and prolong sensors lifetime.

Fig.8 shows the variation of the average number of clusters with respect to the transmission range. The results are shown for varying N. We observe that the average number of clusters decreases with the increase in the transmission range. As we can see in Fig.10, the proposed algorithm produced 16% to 35% fewer clusters than WCA when the transmission range of nodes was 10m. If the node density increased, BS-WCA produced constantly fewer clusters than WCA regardless of node number. When there were 70 nodes in the network, the proposed algorithm produced about 47% to 73% fewer clusters than WCA. According to the result, our algorithm gave better performance in terms of the number of clusters when the node density and transmission range in the network are high.

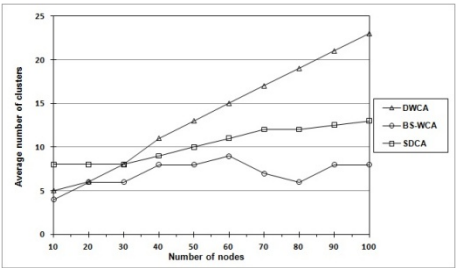


Fig. 7 Average number of clusters vs number of nodes (N) for BS-WCA, DWCA and SDCA

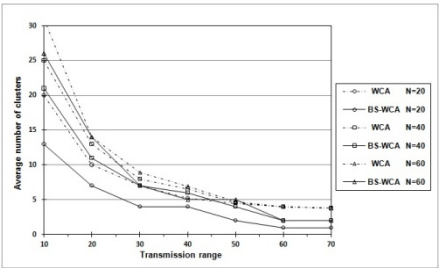


Fig. 8. Average number of clusters vs transmission range BS-WCA and WCA

Fig.9 depicts the average number of re-affiliations that are formed with respect to the total number of nodes in the network. We propose to generate homogeneous clusters whose size lies between two thresholds: $Thresh_{Upper} = 18$ and $Thresh_{Lower} = 9$. The number of re-affiliations increased linearly if there were 30 or more nodes in the network for both WCA and DWCA, but for our algorithm the number of re-affiliations increased starting from 50 nodes. According to the results, our algorithm gave better performance in terms of number of re-affiliations. The main reason is that the frequency of invoking the clustering algorithm is lower in BS-WCA, thus resulting in longer duration of stability of the topology. The benefit of decreasing the number of re-affiliations mainly comes from the localized re-affiliation phase in our algorithm. From Figure 10 it is observed that the sensor nodes 3 and 19 are malicious and have a behavior level less than 0.3. We also note that the sensor 11 is suspicious so if it continues to move frequently it's behavior will gradually be decreased until it reaches the malicious state in this case this node will be deleted from the neighborhood and finally it will be added to the black list. The behavior level of these nodes decreased by 0.001 units when it moves one meter away from its original location but this malicious node does nothing just mobility so in our future works, we will detect the internal misbehavior nodes during distributed monitoring process in WSNs by the follow-up of the messages exchanged between the nodes.

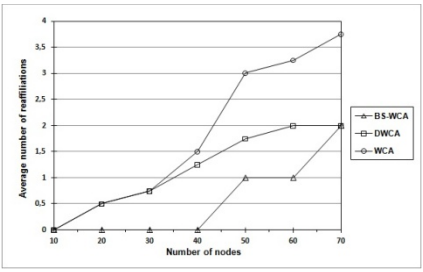


Fig. 9. Average number of re-affiliations

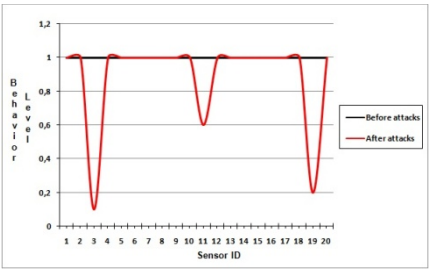


Fig. 10. Behavior level of some sensors before and after attacks

7 Conclusions

In this paper, we have presented a new algorithm called "BS-WCA" for the self-organization of mobile sensor networks. Obtained results from simulations prove that our algorithm outperforms WCA, DWCA and SDCA. It yields a low number of clusters and preserves network structure better than WCA and DWCA by reducing the number of re-affiliations. The proposed algorithm chooses the most robust and safe CHs with the responsibility of monitoring the nodes in their clusters and maintaining clusters locally. As a result of this work, we plan to add a monitoring phase which analyses and detects specific misbehavior in the WSNs by the follow-up of the messages exchanged between the nodes.

Acknowledgements. The authors are grateful to the anonymous referees and Professor Bouhadiba F. for their insightful comments and valuable suggestions, which greatly improved the quality of the paper.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: A Survey. Computer Networks 38(2), 393–422 (2002)
2. Chatterjee, M., Das, S., Turgut, D.: WCA: a weighted clustering algorithm for mobile ad hoc networks. Journal of Cluster Computing (Special Issue on Mobile Ad-hoc Networks) 5, 193–204 (2002)
3. Zabian, A., Ibrahim, A., Al-Kalani, F.: Dynamic Head Cluster Election Algorithm for clustered Ad-hoc Networks. Journal of Computer Science 4(1) (2008)
4. Safa, H., Artail, H., Tabet, D.: A cluster-based trust-aware routing protocol for mobile ad-hoc net-works. Wireless Networks 16(4), 969–984 (2010)
5. Safa, H., Mirza, O., Artail, H.: A Dynamic Energy Efficient Clustering Algorithm for MANETs. In: IEEE International Conference on Wireless & Mobile Computing, Networking & Communication, pp. 51–56 (2008)
6. Elhdhili, M., Azzouz, L., Kamoun, F.: Reputation based clustering algorithm for security management in ad hoc networks with liars. International Journal of Information and Computer Security 3(3), 228–244 (2009)
7. Yu, Y., Zhang, L.: A Secure Clustering Algorithm in Mobile Ad-hoc Networks. In: 2012 IACSIT Hong Kong Conferences, vol. 29, pp. 73–77 (2012)
8. Agarwal, R., Gupta, R., Motwani, M.: Review of Weighted Clustering Algorithms for Mobile Ad-hoc Networks. Computer Science and Telecommunications 33(1), 71–78 (2012)
9. Dahane, A., Berrached, N., Kechar, B.: Energy Efficient and Safe Weighted Clustering Algorithm for Mobile Wireless Sensor Networks. In: The 9th International Conference on Future Networks and Communications, Procedia Computer Science, Niagara Falls, Ontario, Canada, August 17 -20, vol. 34, pp. 63–70 (2014)
10. Soro, S., Heinzelman, W.B.: Cluster head election techniques for coverage preservation in wire-less sensor networks. Ad-Hoc Networks Journal 7(5), 955–972 (2009)
11. Benahmed, K., Merabti, M., Haffaf, H.: Distributed monitoring for misbehavior detection in wireless sensor networks. Security and Communication Networks 6(4), 388–400 (2013)
12. Lehsaini, M., Guyennet, H., Feham, M.: An efficient cluster-based self-organization algorithm for wireless sensor networks. Int. Journal. Sensor Networks 7(1-2), 85–94 (2010)
13. Darabkh, K.A., Ismail, S., Al-Shurman, M.: Performance evaluation of selective and adaptive heads clustering algorithms over wireless sensor networks. Journal of Network and Computer Appli-cations 35(6), 2068–2080 (2012)
14. Choi, W., Woo, M.: A Distributed Weighted Clustering Algorithm for Mobile Ad Hoc Net-works. In: Proc. of the IEEE Advanced International Conference on Telecommunications and Interna-tional Conference on Internet and Web Applications and Services (AICT/ICIW 2006), p. 73 (2006)
15. Abbassi, A., Younis, M.: A Survey on Clustering Algorithms for Wireless Sensor Networks. Computer Communications Journal 30(14-15), 2826–2841 (2007)
16. Chawla, M., Singhai, J., Rana, J.L.: Clustering in Mobile Ad- hoc Networks: A Review. International Journal of Computer Science and Information Security (IJCSIS) 8(2), 293–301 (2010)

17. Er, I.I., Seah, W.K.G.: Mobility-based d-Hop Clustering Algorithm for Mobile Ad Hoc Networks. In: IEEE Wireless Communications and Networking Conference (WCNC 2004), pp. 2359–2364 (2004)
18. Khalil, I., Bagchi, S., Shroff, N.B.: LITEWORP: a lightweight Counter measure for the wormhole attack in multihop wireless networks. In: International Conference on Dependable Systems and Networks, pp. 612–621 (2005)
19. Hsin, M.L.: Self-monitoring of wireless sensor networks. Computer Communications 29(4), 462–476 (2006)
20. Shaikh, R.A., Jameel, H., Lee, S., Y.J., et al.: Trust management problem in distributed wireless sensor networks. In: Proceedings of the 12th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), pp. 411–414 (2006)
21. Hai, T.H., Huhi, E.N., Jo, M.: A lightweight intrusion detection framework for wireless sensor net-works. Wireless Communications and Mobile Computing (Wiley) 10(4), 559–572 (2010)

Distributed Algorithm for Coverage and Connectivity in Wireless Sensor Networks

Abdelkader Khelil and Rachid Beghdad(✉)

Faculty of Sciences, University Abderrahmane Mira of Béjaïa 06000, Béjaia, Algeria
{khalilabdelkader,rachid.beghdad}@gmail.com

Abstract. Even if several algorithms were proposed in the literature to solve the coverage problem in Wireless Sensor Networks (WSNs), they still suffer from some weaknesses. This is the reason why we suggest in this paper, a distributed protocol, called Single Phase Multiple Initiator (SPMI). Its aim is to find Connect Cover Set (CCS) for assuring the coverage and connectivity in WSN. Our idea is based on determining a Connected Dominating Set (CDS) which has a minimum number of necessary and sufficient nodes to guarantee coverage of the area of interested (AI), when WSN model is considered as a graph. The suggested protocol only requires a single phase to construct a CDS in distributed manner without using sensors' location information. Simulation results show that SPMI assures better coverage and connectivity of AI by using fewer active nodes and by inducing very low message overhead, and low energy consumption, when compared with some existing protocols.

Keywords: Wireless Sensor Network (WSN) · Coverage · Connectivity · Distributed Algorithm · Connected Dominating Set (CDS)

1 Introduction

With the recent advances in micro-electronics technologies and wireless communications, a new type of networks has emerged: Wireless Sensor Networks (WSNs). This type of networks includes a large number of devices called sensors deployed over a geographical area to be monitored. A sensor is able to sense, process and transmit data over a wireless communication channel. The applications of WSNs include battlefield surveillance, healthcare, environmental and home monitoring, industrial diagnosis and so on [1].

A fundamental issue in WSNs is the coverage problem [2, 3] that mainly consists in ensuring continuous and effective observation of geographical area while taking into account some constraints, in particular the connectivity of active sensors. The coverage can be considered as a measure of the monitoring quality produced by a sensor network [4].

WSNs are usually dense and redundant (more than 20 nodes/m^3 [5]). So, the coverage of AI can be done, but it is not optimal if all nodes contribute for observing this AI. So, this drawback motivates a connected cover set (CCS) to be employed in a WSN. Conceptually, a CCS is a set of active nodes, which can ensure coverage and connectivity. So, it provides many advantages to QoS of network.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 442–453, 2015.
DOI: 10.1007/978-3-319-19578-0_36

The WSN use the Connected Dominating Set Algorithm to construct a temporary CCS. Only the dominating nodes are responsible for sensing area, and other nodes (dominated nodes) can close the communication modules to save energy, in order to make the network life maximum. Various CDS algorithms [6-12] have been developed but they still suffer from some weaknesses, this is the reason why we focused on the solution of such a problem.

In this paper, we present a novel energy-efficient CDS algorithm for WSN called a Single Phase Multiple Initiator (SPMI). The main contributions of our solution are: (1) high coverage ratio, (2) small number of active nodes, (3) connectivity guaranteed and (4) very low communication overhead, which reduces energy consumption

The rest of this paper is organized as follows. Section 2 presents related work; Section 3 presents concepts relative to graph theory, a set of notations, assumptions of our work and the problem definition. Our solution will be described in Section 4. In Section 5, simulation results will be presented and finally, Section 6 concludes this paper.

2 Related Work

Numerous algorithms for constructing a *CDS* have been surveyed in literature. We cite some of them as follow:

A simple distributed and localized algorithm is proposed in [8] called *CDS-Rule-K* algorithm. It constructs a *CDS* in two phases. The first phase uses marked method to generate a non-optimal *CDS*. Initially all nodes broadcast hello message to receive neighbor tables, and exchange their neighbor tables. If a neighbor node is not covered by other nodes, then it is marked as a node of *CDS*. The second phase uses pruning rules to cut redundant leaf nodes. The pruning rule specifies that if all adjacent nodes are covered by marked brother nodes, then the node is a redundant leaf node, so it is pruned and broadcasts an updates message.

In [9], Yuanyuan and al. present an energy-efficient *CDS* algorithm (*EECDS*), it is based on two phases. In first phase solves a maximal independent set (*MIS*). Initially, all nodes are dyed white. The algorithm started from a white node, while it is dyed black and broadcasts a black message. When receiving a black message, a white neighbor node was stained gray and broadcasts a gray message. When receiving a gray message, a white neighbor node broadcasts query messages to get the states and priorities of nodes around, and sets a timer. If the timer times out ago, it did not receive any black message from its adjacent nodes, then it is dyed black and broadcasts a black message, or remain white until all the nodes in the network were stained gray or black. All the black nodes form a *MIS*. The algorithm in the second phase selects a number of connection nodes to connect the *MIS*. It starts from a non-independent node, while it is dyed blue and broadcast a blue message. When receiving a blue message, an independent node is dyed blue and broadcasts invitation messages. When receiving the invitation message, non-independent nodes compute the priority and broadcast update messages. A non-independent node with the greatest priority is stained blue and broadcasts a blue message until all the black nodes were stained blue. All the blue nodes form a *CDS*.

In [10], Wightman and Labrador have proposed a *CDS* algorithm called *A3*. It uses four forms of messages: Hello message, Children recognition message, Parent recognition message and sleeping message. The sink node starts the protocol by transmitting an initial hello message to their neighboring nodes. Nodes which are not in the range of sink node then this node accepts the message has not been covered by another node; it sets its state as covered, selects the transmitter as its parent node and answers back with a Parent recognition message. If a parent node does not accept any Parent recognition messages from its neighbors, it also turns off. The parent node sets a certain amount of time to accept the answers from its neighboring node. Once this time out, the parent node sorts the list in decreasing order according to the selection metric. Then, parent node broadcasts a children recognition message that includes the complete sorted list to all its candidates. Once the candidate nodes accept the list, they set a timeout period proportional to their position on the candidate list. During that timeout nodes wait for sleeping message from their brothers. If a node accepts a sleeping message during the time out period, it turns itself off.

In [11], Sajjad Rizvi and al. have proposed a *CDS* algorithm called *TC1* for improving the algorithm in [10]. It uses only one type of message: Hello message contains the parent *ID* of the Sender. The initiator node (sink) starts the protocol by transmitting hello message to their neighboring nodes. The neighbor nodes which received a hello message record their parent *ID* and calculate a timeout period according to their residual energy and distance from the sender. The child node which expires its timeout sends a hello message to its neighboring nodes too. So if the parent node receives this message, it will be a dominator node. This process continues until the complete topology is formed with nodes acting as *CDS* for rest of the nodes in the network.

In [12], ShiTing-jun and al. have proposed a *CDS* algorithm called *IPCDS*. It uses the staining and markers methods to solve the *MIS* and *CDS*, and uses the pruning rule to further reduce the *CDS*. Initially, all nodes are dyed white and have been marked. The initiator node (white node) starts the protocol; it is dyed black and broadcasts a black message. When at first receiving black, a white neighbor node is marked as the child of nodes broadcasting message. Upon receiving a black message, if the white neighbor node is not marked, it is dyed gray and broadcasts gray messages. Upon receiving a gray message, if the white neighbor node is not marked, then according the residual energy and *RSSI* (Received Signal Strength Indicator) sets the timer value. If the timer times out ago, it received a black message by broadcasted the brother node, it is dyed gray and broadcasts gray messages, or it is dyed black and broadcasts black messages. Upon receiving a black message broadcasted by the child node, the gray node was stained black and broadcasts black messages. Upon the black node is in line with the pruning rule, it is dyed gray and broadcasts gray messages. This process continues until all nodes in the network are stained gray or black. At the end of algorithm, all the black nodes form a *CDS*.

3 Preliminaries

A. WSN Model

The network is modeled by a graph $G = (S, E)$, where S represents the vertices set and E the set of edges. An edge between two vertices u and v exists if u can communicate

v. For a sensor *u*, it characterizes by alone identity denoted *ID(u)* and we distinguish two different ranges: communication range, denoted *CR*, and sensing range, denoted by *SR*. Two sensors are communicated if and only if the distance between them is at most equal to *CR*. The covered area from a node u (also called monitored or sensed area) is the surface within which if an event occurs it will be sensed by the sensor u. This area is modeled as a disk of radius *SR* centered at *u*. Similarly, the communication area, inside which the sensor u can send and receive messages, is modeled by a circle of radius *CR* centered at *u*. In this work, we consider $CR \leq 2 * SR$.

B. Definitions

In this subsection we define the concept on which our work is based, i.e. Connected Dominating Set (*CDS*).

- Connected Dominating Set:

Given an undirected graph *G=(S,E)*, a Dominating Set (*DS*) of *G* is a subset of vertices $D \subseteq S$, such that any vertex *u* of the graph is either in *D* or has a neighbor $v \in D$ [18]. A graph has more than one dominating set. When a *DS* is connected, it is denoted as a *CDS*; that is, any two nodes in the *DS* can be connected through intermediate nodes from the *DS*.

C. Solution Assumptions

In this work, we assume a randomly deployed network. Once disseminated, sensors are assumed to be static. The network consists of nodes deployed in high density in order to ensure initial connectivity. Furthermore, the network is homogeneous, that is all sensors have the same sensing radius and the same communication ranges. We also assume that sensors have a unique identifier (*ID*). Finally, we assume that each sensor knows its degree.

D. Problem Definition

The random deployment of sensors is the most used for a broad range of applications in inaccessible environments. Due to the unplanned nature of this deployment type, a *WSN* could lead to sensing holes which decrease drastically the reliability of the network. In order to overcome this shortcoming, sensor nodes are disseminated in high density. Although, dense deployment minimizes the sensing holes, allows fault tolerance and increases the reliability of applications, it has its own drawbacks; maximizes the redundancy which decreases energy efficiency. Monitoring the same region of the interest area by several sensors involves a waste of energy. This behavior is in conflict with the most critical constraint of a *WSN* (energy efficient). Thus, it is crucial to have a solution that reduces redundancy in order to assure a good coverage ratio and connectivity.

4 SPMI Solution

A. SPMI Overview

The *SPMI* requires a single phase to generate a *CDS*. Nodes in the *CDS* are called dominators while the *nonCDS* nodes are referred to dominated. The aim of the *SPMI*

is to generate a small set of dominators while keeping the message overhead and energy consumption low. The suggested algorithm allows the construction of *CDS* which has a minimum number of necessary and sufficient nodes in a distributed manner. In fact, each sensor performs the algorithm independently from the others in order to determine its status: *Dominator* (active) or *Dominated* (passive). Initially, all sensors are in uncovered state for a timeout and make their decision to be in active or in passive state. The active nodes form a *CDS* of the network, they provide coverage (monitoring) of the interest area.

The nodes having a higher level energy than their one hop neighbors, they will be the parents of their neighbors and they form *DS*. To make this set connected, we activate the child nodes that have a higher degree and further away from their parents. So the brothers of active child node will be in passive state.

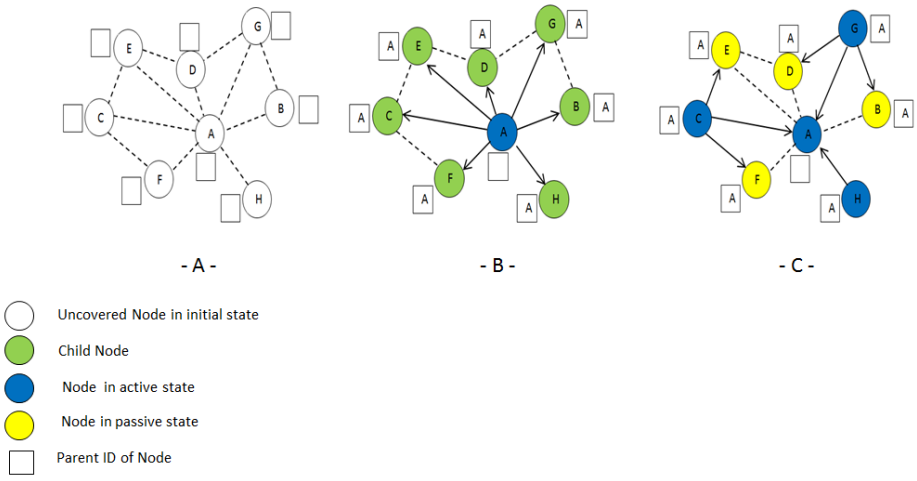


Fig. 1. An illustrative example

Figure 1.A represents a simple topology of 08 nodes which are uncovered in initial state; each node computes a timeout based on its level energy. In the figure 1.B, the node *A* will be in active state because it has a higher level energy than its neighbors, and it broadcasts a message which is received by nodes (*B, C, D, E, F, G* and *H*) under its communication area. After receiving, they will be children of node *A* and they recalculate (update) their timeout based on the degree and further away from their parents.

In the figure 1.C; When the timeout of nodes *C*, *G* and *H* expire without receiving any message from other nodes, they turn themselves in active state and broadcast a message (including their *ID* and parents *ID*) to their neighbors. The children nodes *E* and *F* are turned themselves in passive state after receiving a message which has the same parent *ID* from the node *C*. Also; the children nodes *B* and *D* are turned themselves in passive state after receiving a message from the node *G*.

So our algorithm maintains the node *A* is in active state, and in its communication range, it actives only three nodes (*C, G and H*) to assure more the coverage and

connectivity; and put the other nodes in passive state which never send message to keep their energy.

B. SPMI Description

Each sensor *i* is characterized by variables:

E_{ini}: initial energy (assumed to be the same for all nodes).

RE_i: residual energy.

$REP_i=E_{ini}/RE_i$: residual energy percentage.

T_{cons}: time constant.

T_i: waiting time or timeout.

ID_i: identifier of the node *i*.

$Parent_iID$: parent identifier of the node *i*.

$State_i$: indicates the sensor state, it may take one of values: *Uncovered, Dominator or Dominated.*

$Degree_i$: the one-hop neighbors number of node *i*.

$RSSI_j$: the signal strength of parent node *j* received by the child node *i*. It uses for estimation the distance between the nodes [16] [17].

$RSSI_c$: the minimum required signal strength to ensure connectivity.

The functions used by a node *i* are:

receive msg (ID_j, $parent_jID$): that is the node *i* which has received a domination message from an active neighbor *j*.

send msg (ID_i, $parent_iID$): that is the node *i* which has sent a domination message.

calculate (T_i): the node *i* computes a timeout T_i according to the formula (1). A timeout is inversely proportional to the remaining energy level.

$$T_i = T_{cons} / REP_i \tag{1}$$

Recalculate (T_i): when the node *i* receives a message from its parent node *j*, then the node *i* recalculates the timeout according to the formula (2). A timeout inversely proportional to the remaining energy level, Degree and distance between the node *i* and node *j*.

$$T_i = T_{cons} / (REP_i + Degree_i + (RSSI_c / RSSI_j)) \tag{2}$$

So, at first each node *i* computes a timeout *Ti* according to the formula (1) (*line1*). Sensors with a higher residual energy percentage have a shorter timeout that expires earlier. Therefore, these sensors have more chance to be in active state. Sensors with a lower residual energy percentage have a longer timeout that expires later.

During this time, the sensor listens to messages sent by neighbors (*lines 2&4*): When the first receiving the message (*line7*), then the node will be the child of node broadcasting message and it recalculates its timeout according to the formula (2) (*line8*) (sensors with a higher residual energy percentage, higher degree and farther from parent node have a shorter timeout that expires earlier). If the node receives another message and its parent *ID* is the same of the parent *ID* in the message received (*line11*), then the node decides directly to change its state to a *Dominated* without sending any message (*line16*). If the timeout expires without receiving any message or receiving massages having parent *ID* different from the node's parent *ID*, the node then concludes that it is *Dominator* node (*line19*) and broadcasts a message announcing domination to its one-hop neighbors (*line20*). At the end of the algorithm, all *Dominator* nodes are members of the *CDS*.

SPMI algorithm is formally as follows:

For all Sensor *i*

BEGIN
1. *Calculate(T_i); $parent_i$ID=void; $State_i$=Uncovered; Verf=true*;
2. While (*T_i <> 0)* do
3. Begin
4. Listen;
5. If *receive msg (ID_j, $parent_j$ID)* then
6. Begin
7. If *($parent_i$ID is void)* then
8. *$parent_i$ID =ID_j ; recalculate(T_i) ;*
9. Else
10. Begin
11. If *($parent_i$ID == $parent_j$ID)* then
12. *Verf=false*; Break;
13. End if;
14. End if;
15. End While;
16. If (*Verf == false)* then *$State_i$ = passif (dominated)*
17. Else
18. Begin
19. *$State_i$ = active ; (dominator)*
20. *Send msg (ID_i, $parent_i$ ID) ;*
21. End if ;
END.

Figure 2 will illustrate the state diagram of the *SPMI* algorithm

Case1*: if the parent ID of node is different from the parent ID in any received message or the node never receives any message.*

Case2*: if the parent ID in the received message is the same of the node's parent ID.*

Fig. 2. The state diagram of the *SPMI* algorithm

5 Simulation

We simulate our solution by using Java language to evaluate *SPMI* Algorithm and compare its performance to other Algorithms.

A. Simulation Parameters

Experimental results were obtained from randomly generated networks in which nodes are deployed over a square sensing field. The initial graph, the one formed right after the deployment, is connected. Simulations were carried over densities varying from 10 to 100 nodes. The results presented hereafter are the average of 100 iterations for each simulated scenario. The performance metrics include: 1- number of active nodes; 2- number of messages used in the CDS building process; 3- amount of energy used in the process; and 4- coverage ratio. Table 1 lists all the parameters used in simulation.

Table 1. Simulation parameters

Parameter	Value	Parameter	Value
Range	200mx200m	*SR*	50 m
Nodes	10,20,40,60,80,100	$RSSI_c$	80
CR	63m	T_{cons}	100 ms

B. Performance Evaluation

In this subsection, we compare the performances of *SPMI* with other solutions: *A3* algorithm [10]; *EECDS* [9]; *CDS-Rule-K* algorithm [8]; *TC1* algorithm [11] and the *IPCDS* algorithm [12].

1- CDS Size :

Figure 3 shows that when the network density increases, the numbers of nodes generated by the six kinds of algorithm are increased. It is clear that *SPMI* generated less *CDS* size compared to *EECDS* and *CDS-Rule-k*; and it is nearly similar to *IPCDS*. But it is more than *TC1* and *A3*. This difference of active nodes size is exploited by *SPMI* for assuring more connectivity and coverage in *WSN*, such that only 6 to 17 nodes are active for different sensors populations.

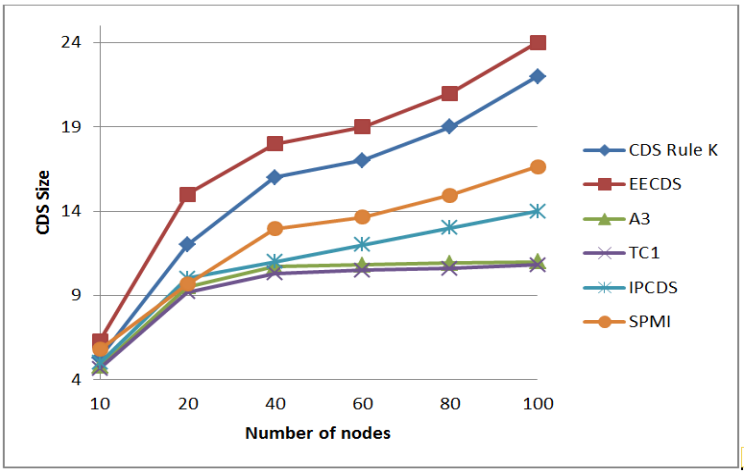


Fig. 3. CDS size versus network size

2- Message Overhead:

Figure 4 shows the message overhead of the six kinds of algorithm with respect to network size. The message overhead was evaluated based on the number of messages sent by nodes during the *CDS* construction. *SPMI* requires significantly lower message overhead compared to all five algorithms when the network density increases.

The efficiency of *SPMI* algorithm in terms of the number of message sent is due to it requires a single phase and one message at most for each node. Contrary to *EECDS,* CDS- *Rule-k* and *IPCDS* which require two phases and high amount of exchanges of messages, the *A3* and *TC1* need a single phase but the number of message is high than *SPMI*, they require three messages and one message respectively for each node.

3- Energy Consumption:

In order to evaluate the energy efficiency, we used a discrete energy model. Every node has an initial energy equal to 100 units. An active node consumes 1 unit of energy during 1 unit of time and 0 if it is passive. The energy required to transmit a message is 1 unit and the one spent for its reception is 0.2; the consumption in listening state is the same one as at the reception of a message. Notice that these energy consumptions are in correspondence with the reality. Indeed, for a Mica2 sensor [14], the energy spent in listening state is equal to that required for the reception of a message and energy used to transmit a message is equal to five times the energy of its reception [13, 15].

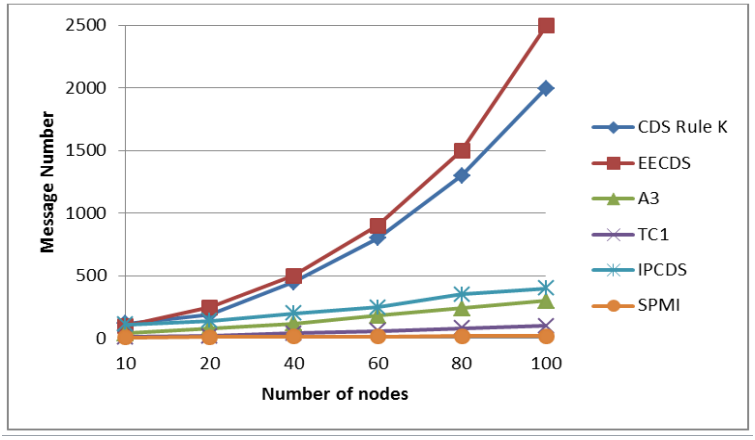


Fig. 4. Message overhead versus network size

Figure 5 represents the total energy consumption by the all six algorithms while varying the deployed nodes number.it is clear that *EECDS* and *CDS-Rule-k* algorithms consume a high significant amount of energy and their energy increases linearly with the number of neighbors. The other algorithms consume less amount of energy; they are similar and their energies are nearly constant with the size of network; but the *SPMI* consumes the lowest energy for constructing the *CDS*. This can be explained by the low number of messages exchanged between nodes. This shows that our algorithm is scalable and can be used for a large network deployment.

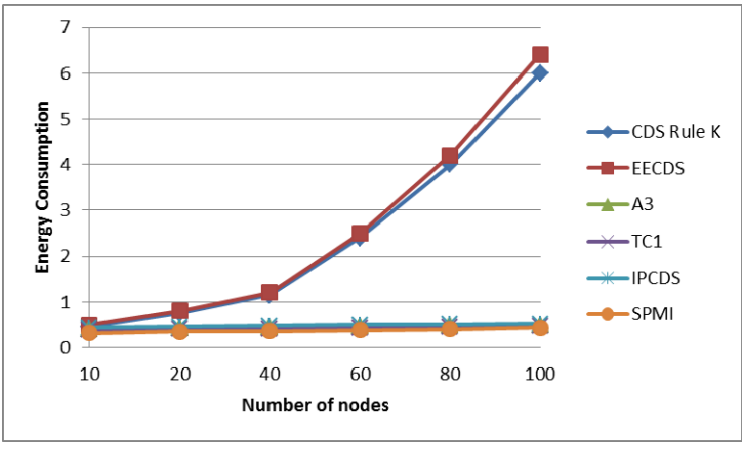


Fig. 5. Total energy consumption versus network size

4- Coverage Ratio:

The coverage ratio is evaluated by dividing the deployment area to cells. A cell is considered covered if its center is covered.

Figure 6 represents the average coverage ratio which is defined as the percentage of interest area covered by active nodes of the four algorithms.

We can say that although the two algorithms (*EECDS* and *CDS-Rule-k*) produce an almost similar coverage with the selected active nodes. The *A3* algorithm covers the same or more area compared to *EECDS* and *CDS-Rule-k*.

SPMI is still better; it covers more area than other algorithms. Such as it provides a better coverage ratio with 86.29% for the lowest density, this ratio increases gradually until it exceeds 99.66% for the highest density. For 100 deployed nodes, it is shown in Figure.5 that *SPMI* provides an improvement of coverage ratio equal to 3.96%, 2% and 1.96% compared to *CDS-Rule-k*, *EECDS* and *A3* respectively. This is due to select far nodes from the parent node according to formula (2).

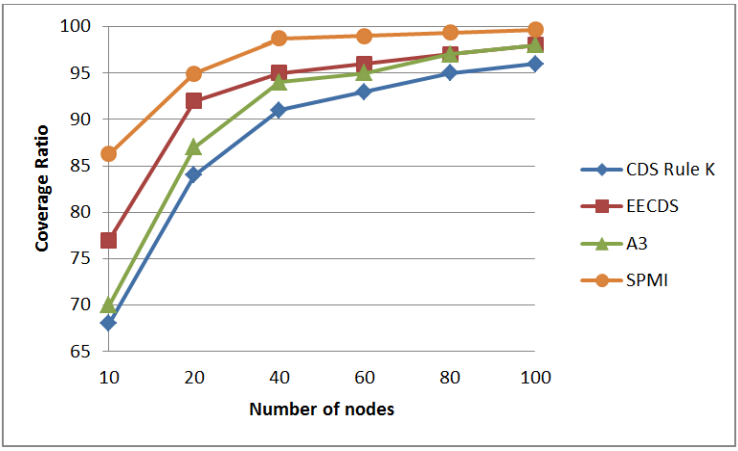


Fig. 6. Coverage ratio versus network size

6 Conclusion

In this paper, we have proposed a distributed algorithm called *SPMI* that can construct a *CDS* in a single phase to maintain the coverage and connectivity in Wireless Sensor Networks. The *SPMI* limits the number of exchanged messages among nodes and keeps the number of active nodes low. Simulation has been done to validate the effectiveness of the suggested algorithm. The results show that, *SPMI* outperforms the other algorithms [8-12] in terms of coverage ratio which is the most important metric. It also competes perfectly in terms of selected active nodes while reducing the communication overhead significantly, what decreases the energy consumption.

Our future work will focus the coverage and connectivity problem in case of mobile nodes and with the presence of obstacles.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks Journal 38(4), 393–422 (2002)
2. Huang, C.F., Tseng, Y.C.: A survey of solutions to the coverage problems in wireless sensor networks. Journal of Internet Technology 6(1), 1–8 (2005)
3. Cardei, M., Wu, J.: Energy-efficient coverage problems in wireless ad hoc sensor networks. Computer Communications Journal 29(4), 413–420 (2006)
4. Meguerdichian, S., Koushanfar, F., Potkonjak, M., Srivastava, M.B.: Coverage problems in wireless ad-hoc sensor networks. In: 20th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 1380–1387 (2001)
5. Rajavavivarme, V., Yang, Y., Yang, T.: An overview of wireless sensor network and applications. In: Proceedings of the 35th Southeastern Symposium on System Theory, pp. 432–436 (March 2003)
6. Khelil, A., Beghdad, R.: Coverage and Connectivity Protocol for Wireless Sensor Networks. In: Proceedings of The 24th International Conference of Microelectronics ICM 2012, Algeria, December 17-20 (2012)
7. Pazand, B., Datta, A.: Minimum dominating sets for solving the coverage problem in wireless sensor networks. In: Youn, H.Y., Kim, M., Morikawa, H. (eds.) UCS 2006. LNCS, vol. 4239, pp. 454–466. Springer, Heidelberg (2006)
8. Wu, J., Cardei, M., Dai, F., Yang, S.: Extended dominating set and its applications in ad hoc networks using cooperative communication. IEEE Trans.on Parallel and Distributed Systems 17(8), 851–864 (2006)
9. Yuanyuan, Z., Jia, X., Yanxiang, H.: Energy efficient distributed connected dominating sets construction in wireless sensor networks. In: Proceedings of the ACM International Conference on Communications and Mobile Computing, pp. 797–802 (2006)
10. Wightman, P.M., Labrador, M.A.: A3: A Topology Construction Algorithm for Wireless Sensor Network. In: Proc. IEEE Globecom (2008)
11. Karthikeyan, A., et al.: Topology Control Algorithm for Better Sensing Coverage with Connectivity in WSN. Journal of Theoretical and Applied Information Technology JATIT (June 2013)
12. Shi, T., Shi, X., Fang, X.: A Virtual Backbone Construction Algorithm Based on Connected Dominating Set in Wireless Sensor Networks. In: Proceedings of the 2014 International Conference on Computer, Communications and Information Technology (CCIT) (2014)
13. Ye, F., Zhang, H., Lu, S., Zhang, L., Hou, J.: A randomized energy-conservation protocol for resilient sensor networks. Wireless Networks 12(5), 637–652 (2006)
14. MICA2 Mote Datasheet. Available from Crossbow Technology Inc. (2009), `http://www.xbow.com/`
15. Anastasi, G., Falchi, A., Passarella, A., Conti, M., Gregori, E.: Performance measurements of motes sensor networks. In: Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 174–181 (2004)
16. Pu, C.-C., Chung, W.-Y.: Mitigation of Multipath Fading Effects to Improve Indoor RSSI Performance. IEEE Sensors Journal 8(11), 1884–1886 (2008)
17. Hood, B., Barooah, P.: Estimating DoA From Radio-Frequency RSSI Measurements Using an Actuated Reflector. IEEE Sensors Journal 11(2), 413–417 (2011)

Optimizing Deployment Cost in Camera-Based Wireless Sensor Networks

Mehdi Rouan Serik(✉) and Mejdi Kaddour

LITIO Laboratory, University of Oran 1, BP 1524, El-M'Naouer, 31000 Oran, Algeria
{rouan.mehdi,kaddour.mejdi}@univ-oran.dz

Abstract. We discuss in this paper a deployment optimization problem in camera-based wireless sensor networks. In particular, we propose a mathematical model to solve the problem of minimizing the number of cameras required to cover a set of targets with a given level of quality. Since solving this kind of problems with exact methods is computationally expensive, we rather rely on an adapted version of *Binary Particle Swarm Optimization* (BPSO). Our preliminary results are motivating since we obtain near-optimal solutions in few iterations of the algorithm. We discuss also the relevance of hybrid meta-heuristics and parallel algorithms in this context.

Keywords: Camera-based wireless sensor networks · Minimum cost deployment · Coverage quality · Binary particle swarm optimization

1 Introduction

Wireless Sensor Networks (WSN) are particular ad-hoc networks defined as a set of cooperating nodes disseminated in a given geographic area in order to collect its data about some phenomenon autonomously. Specifically, Camera-based wireless sensor networks (WSN) form an emerging research area with many promising applications. Potential applications include remote video surveillance, monitoring and assisting elderly and health patients, and habitat monitoring.

We study in this paper the cost deployment of camera-based WSNs, where the main concern is to determine the optimal minimum number of cameras, along with their positions and their orientations to track a given set of targets with a prescribed level of quality. This problem can be solved with various exact mathematical programming tools such as Branch and Bound. However, as the problem size increases, solving such problems using these exact methods becomes computationally intractable. In fact, Ai and Abouzeid [1] have demonstrated that this problem is NP-hard. A traditional way to deal with such difficult problems is to rely upon meta-heuristic methods. In particular, we adapt the generic procedure of *Particle Swarm Optimization* (PSO) to solve this deployment problem. We also introduce a quality coverage parameter which serves to ensure that each target in the solution is covered with a sufficient level of quality. Indeed, as the targeting objects move away from the sensing camera, the level of details falls off. Our preliminary results show that the proposed method behaves well both in computational performances and solution quality.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 454–464, 2015.
DOI: 10.1007/978-3-319-19578-0_37

The rest of paper is organized as follows. Section 2 reviews the relevant literature on this topic. Section 3 formulates the problem and describes the proposed coverage model. The detailed mathematical model is given in Section 4, while the proposed PSO-based algorithm is described in Section 5. Section 6 assesses and discusses the obtained results. Finally, The last section concludes the paper and suggests some future research directions.

2 Related Works

Existing literature in the field distinguishes typically two important kinds of deployment problems: target coverage and area coverage. We focus here on the first one. One version of the problem consists on covering a maximum number of targets with a minimum number of sensors. In [1], authors proposed an exact integer linear program (ILP) and a Centralized Greedy Algorithm (CGA) for the maximum coverage with minimum sensors (MCMS) problem. Then, they provided a Distributed Greedy Algorithm (DGA) solution. They showed that DGA does better than the two other methods (ILP and CGA) by incorporating a measure of the sensors' residual energy into DGA. Aziz *et. al.* [2] proposed a new algorithm to optimize sensor coverage using PSO and Voronoi diagrams. PSO is used to find the optimal deployment of sensors providing the best coverage, while Voronoi diagram is used to evaluate the fitness of the solution. They showed that the proposed algorithm achieves a good coverage with a better time efficiency than existing approaches.

Authors in [12] improved the field of view (FOV) coverage of a camera network. They considered randomly scattered cameras in a wide area, where each camera may adjust only its orientation and not its localisation. They also implemented a PSO algorithm and efficiently found an optimal orientation for each camera. They considered also region of interest in the search space (ROI) and occlusions. In [6], authors considered the area coverage problem in a 2D/3D-grid space. The solution process was based on Particle Swarm Optimization Inspired Probability (PSO-IP). For comparison purposes, they also implemented alternative methods based on Tabu Search, genetic algorithms and simulated annealing. Results showed that the proposed PSO-IP overcomes the three other methods especially with large instances. Unlike these two works, we deal with targets coverage not area coverage. We give also a mathematical model to the coverage problem. Note that our solution approach is partly inspired by this last work. But as far as we know, no existing work has formulated a problem similar to ours, in particular by considering a continuous measure of sensing quality.

3 Problem Definition

We assume that a set of N camera sensors $\{S_i : i = 1, \ldots, n\}$ are deployed on the euclidean space A. For each camera S_i, we are given its Cartesian coordinates (X_{S_i}, Y_{S_i}) and its orientation φ in A. The field of vision of each camera is modeled

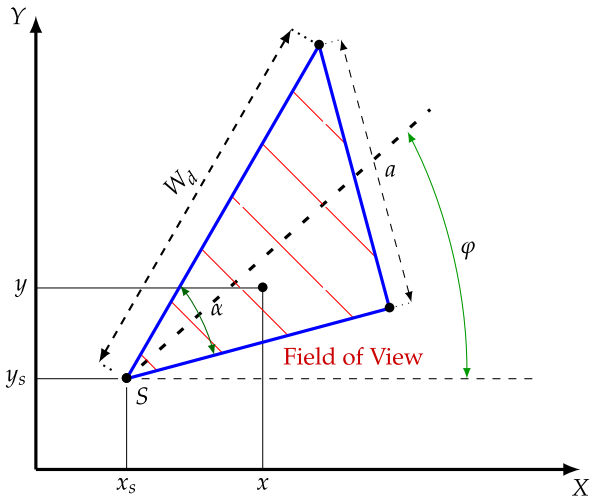


Fig. 1. Camera's field of view

as in [8] by isosceles triangle as depicted in Fig. 1, where α represents the aperture of the camera and W_d its working distance.

A target located on the coordinates (x, y) is assumed to be covered by a given camera with coordinates (x_s, y_s) if the three following constraints are satisfied:

$$cos(\varphi) \cdot (x - x_s) + sin(\varphi) \cdot (y - y_s) \leq W_d \tag{1a}$$

$$-sin(\varphi) \cdot (x - x_s) + cos(\varphi) \cdot (y - y_s) \leq \frac{a}{2W_d} \cdot (cos(\varphi) \cdot (x - x_s) + sin(\varphi) \cdot (y - y_s)) \tag{1b}$$

$$-sin(\varphi) \cdot (x - x_s) + cos(\varphi) \cdot (y - y_s) \geq -\frac{a}{2W_d} \cdot (cos(\varphi) \cdot (x - x_s) + sin(\varphi) \cdot (y - y_s)) \tag{1c}$$

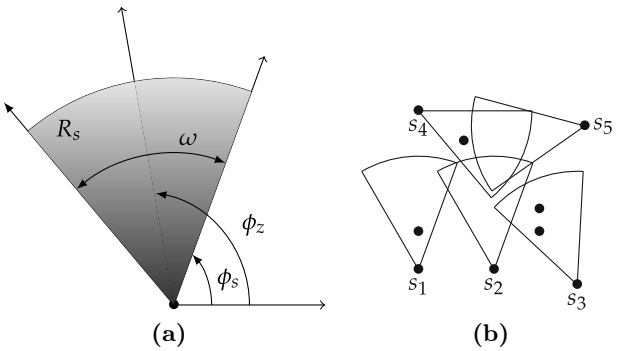
3.1 Coverage Model

Coverage models determine first if a given target can be covered or not by some sensor, but can also measure a corresponding quality parameter [11]. This is accomplished by calculating the geometric relation between sensors and targets. In most cases, it consists in calculating the euclidean distance and angles, but some research works assume also that the sensing quality of a sensor is reduced with the increase of the distance away from the sensor [5], [10]. In our case, we make a similar assumption by adopting a directional disk model where the

coverage parameter for a given sensor/target pair is represented by a non negative real number calculated as follows:

$$f(d(s,z)) = \frac{C}{d^{\alpha}(s,z)} \tag{2}$$

where $d(s,z)$ is the distance separating camera s from target z, α is the exponent attenuation and C is a constant. In particular, we assume that quality decreases quadratically as a function of the distance ($\alpha = 2$).



(a) **(b)**

Fig. 2. Illustrations of : 2a Directional Model ; 2b Three active cameras covering 4 targets

4 Mathematical Model

The adopted model assumes a known number of targets, in a given space, where the objective is to cover these targets optimally. Table 1 define the formal notations used in our model.

Table 1. Problem's formal notations

Variables	**Designation**
n	number of possible camera locations
m	number of possible angles
t	number of targets
δ	quality parameter
d_{ij}	distance between camera i and target j
(x_i, y_i)	Cartesian coordinate of the camera i
(x_j^t, y_j^t)	Cartesian coordinates of the target j
φ_k	angle k

let b_{ij} and O_{ik} binary variables defined as follows:

$$b_{ij} = \begin{cases} 1 & \text{if camera at position } i \text{ covers target } j \\ 0 & \text{otherwise} \end{cases}$$

$$O_{ik} = \begin{cases} 1 & \text{if camera at position } i \text{ has active angle } k \\ 0 & \text{otherwise} \end{cases}$$

The objective of our optimization problem is defined as follows:

$$\min \sum_{i=1}^{n} \sum_{k=1}^{m} O_{ik} \tag{3}$$

This objective implies the minimization of the number of deployed cameras, provided that the following constraints are satisfied.

$$\sum_{k=1}^{m} O_{ik} \leq 1, \quad \forall i = 1, \ldots, n. \tag{4}$$

The above constraint ensures that at most one angle is active per each camera. $\sum_{k=1}^{m} O_{ik} = 0$, corresponds to the situation where no camera is deployed at location i.

$$\sum_{i=1}^{n} b_{ij} \geq 1, \quad \forall j = 1, \ldots, t. \tag{5}$$

This constraints ensures that each target is covered by at least one camera.

Now, the minimum coverage quality for each target is satisfied through the following constraint:

$$\sum_{i=1}^{n} b_{ij} \frac{C}{d_{ij}^{\alpha}} \geq \delta, \quad \forall j = 1, \ldots, t. \tag{6}$$

where C is a constant and α is the attenuation exponent.

Given some locations of a camera i oriented with angle k, the following three constraints enforce, as described above in (1a),(1b) and (1c), that some target j is covered properly if $b_{ij} = 1$.

$$\sum_{k=1}^{m} O_{ik} \left[cos(\varphi_k) \cdot (x_j^t - x_i) + sin(\varphi_k) \cdot (y_j^t - y_i) \right] \leq d + L_1(1 - b_{ij}),$$
$$\forall i = 1, \ldots, n, \quad \forall j = 1, \ldots, t. \tag{7}$$

$$\sum_{k=1}^{m} O_{ik} \left[- \left(sin(\varphi_k) + \frac{a}{2d} cos(\varphi_k) \right) (x_j^t - x_i) + \left(cos(\varphi_k) - \frac{a}{2d} sin(\varphi_k) \right) (y_j^t - y_i) \right]$$
$$\leq L_2(1 - b_{ij}) \quad \forall i = 1, \ldots, n, \quad \forall j = 1, \ldots, t. \tag{8}$$

$$\sum_{k=1}^{m} O_{ik}\left[\left(sin(\varphi_k) - \frac{a}{2d}cos(\varphi_k)\right)(x_j^t - x_i) - \left(cos(\varphi_k) - \frac{a}{2d}sin(\varphi_k)\right)(y_j^t - y_i)\right] \leq L_3(1 - b_{ij}) \quad \forall i = 1, \ldots, n, \quad \forall j = 1, \ldots, t. \tag{9}$$

where L_1, L_2, L_3 are large constants, which are introduced to make theses constraints meaningless when $b_{ij} = 1$.

The above objective with the introduced constraints define a mixed-integer problem which is hard to solve in general. In particular, it is not easy to move from one solution to another when applying some meta-heuristic algorithm for example. Hence, we relax this model by moving the coverage constraints (5)-(6) into the objective, as follows.

$$\min \left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} O_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{t} b_{ij}\left(1 + \frac{C}{d_{ij}^{\alpha}}\right) \right\} \tag{10}$$

Subject to (4),(7)-(9).

5 Particle Swarm Optimization

Particle swarm optimization (PSO) is a meta-heuristic method invented by *Russel Eberhart* (Electrical Engineer) and *James Kennedy* (socio-psychologist) in 1995. This algorithm, inspired by social behaviour, has been introduced as an optimization tool dealing with real numbers initially and with integers lately [7]. It is mostly inspired from the manner in which a flock of birds moves with various individuals leading the flock during the travel at different periods of time. The PSO algorithm consists of a group of individuals named particles. Each particle $p <$ `Swarm_size` is a potential solution to an optimization problem, having its own position in the space search. After each iteration, it moves in function of one of its components:

- Actual velocity V ;
- Best solution L_b ;
- Actual position X ;
- Best neighbourhood solution G_{best} ;

The movement of each particle obeys to the equations:

$$V_{k+1} = \omega.V_k + C_1 r_1(P_b - X_k) + C_2 r_2(P_g - X_k) \tag{11}$$

$$X_{k+1} = X_k + V_{k+1} \tag{12}$$

where $X = (x_{ik}), V = (v_{ik}), i = 1, 2, \ldots, n, k = 1, 2, \ldots m$, denote the distance/angle and the velocity vectors, respectively. ω is the inertia weight, r_1 and r_2 are two random numbers uniformly chosen in $[0, 1]$, and C_1, C_2 are constant values. Finally k is the iteration index.

A given solution for the deployment problem is a number of cameras with corresponding positions and active angles (see Fig. 3a). We adopt a binary representation of the position vector X. Besides, since PSOs require movements, we define two different types of moves:

1. Rotation: selecting a different camera for an active camera.
2. Displacement: moving one camera from position i to position i'.

These movements are guided by the velocity parameter. Thus, we redefine (11) as follows:

$$V = (v_{ik}, \forall k = 1, 2, \ldots, m) = \begin{cases} 1 & \text{if } alea > \frac{1}{1+(G_{best}-L_b)} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Where `alea` is a random number from $[0, 1]$, then Eq. (12) become:

$$X_{k+1} = X_k \oplus V_{k+1} \tag{14}$$

A camera rotation is defined through the logical operator "exclusive `OR`", i.e., to invert a zero bit to one in the X vector, while a camera displacement is defined through a binary shift, and this will be a position swap. As a stopping criteria, we define a maximum number of iterations. Algorithm 1 gives an overview of our implemented PSO.

Algorithm 1. Proposed PSO algorithm

for all ($p <$ `Swarm_size`) **do**
  `Random_init(p)`;
  $L_b(p) =$ `Fitness(p)`;
end for
$G_{best} = \min_p \{L_b(p)\}$;
repeat
  `alea=rand()`;
  for all ($p <$ `Swarm_size`) **do**
    Calculate $V(p)$ using (13);
    `move(p)`;
    if `Fitness(p)` $< L_b(p)$ **then**
      $L_b(p) =$ `Fitness(p)`;
    end if
  end for
  $G_{best} = \min_p \{L_b(p)\}$;
  Update positions with (14);
until Stop criteria satisfied

For comparison purposes, we implemented a standard *Simulated Annealing Algorithm* (SA) [9]. The basic idea behind comes from the principles of statistical mechanics whereby the annealing process requires heating and then slowly

cooling a substance to obtain a strong crystalline structure. At each iteration, a random neighbour is generated. Movements that improve the cost function are always accepted. Otherwise, the neighbour is selected with a given probability that depends on the current temperature and the amount of degradation ΔE of the objective function. This probability is calculated as follows:

$$P(\Delta E, T) = e^{\frac{-\Delta E}{T}}$$

where ΔE represents the fitness difference. This is algorithm is used in large-scale optimization problems in wireless sensors networks as in [6], [4].

6 Experiments and Results

We discuss in this section various experiments related to our approach. First, the most important parameter to be defined in order to implement a PSO algorithm is the solution coding or representation. Each solution is encoded in binary where the vector X represents camera positions and angles (Fig 3a) and (Fig 3b). The experiments were executed on a computer with Intel © Core ™ i3-2350 M CPU @ 2.30 GHz CPU and 4.0 GB of RAM.

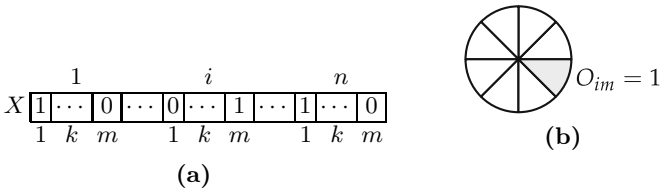


Fig. 3. PSO solution: 3a X Position vector; 3b Active angle

Table 2a gives parameters used to implement the method. `Swarm_size` and `Steps` represent the number of particles (potential solutions) involved in PSO and iteration's number, respectively. C_1 and C_2 represent learning factors of the algorithm, most often set to 2.0 [3]. $\mathtt{W_d}$ represents the working distance of a camera (extrinsic parameter), $\mathtt{X_{max}}$ and $\mathtt{Y_{max}}$ are the grid dimensions. $\mathtt{Q_{max}}$ and $\mathtt{Q_{min}}$ are, respectively, the maximum and the minimum quality coverage of a camera. Here we require a certain level of quality in target covering. C is a constant which specifies how the quality coverage decreases when moving away from the sensor.

In Fig 4, we give an initial solution then a final one found by PSO algorithm. The figures show that we can easily find a good solution consisting of two cameras covering nine targets. As shown in Tab 2b, finding such a solution does not require more than a hundred of iterations in a very brief time for a small number of targets (less than a half of a second). We observe also the high success rate over all executions.

As shown in (Fig 4d) we clearly obtanin better average fitness by the proposed PSO than the one obtained by the SA for fifty executions of the two methods.

Table 2. PSO and SA parameters and executions

(a)

Parameters	Values
`Swarm_size`	30
`Steps`	300
C_1, C_2	1.4
n	10
m	8
t	10
L_1, L_2, L_3	`Max_DBL`
`C`	10.0
$\texttt{Q}_{\texttt{max}}$	0.4
$\texttt{Q}_{\texttt{min}}$	0.05
$\texttt{W}_{\texttt{d}}$	10
$\texttt{X}_{\texttt{max}}, \texttt{Y}_{\texttt{max}}$	100
`T`	1000.0

(b)

Exec	Fit.	n_s	n_t	Steps	time (s)
1	−7.018	2	9	200	0.706
2	−7.050	3	10	167	0.566
3	−8.057	2	10	85	0.109
4	−7.033	2	9	52	0.041
5	−7.048	3	10	1	0.003
6	−7.036	2	9	269	0.78
7	−7.032	2	9	61	0.052
8	−7.035	2	9	170	0.12
9	−7.058	3	10	131	0.509
10	−7.058	3	10	61	0.05

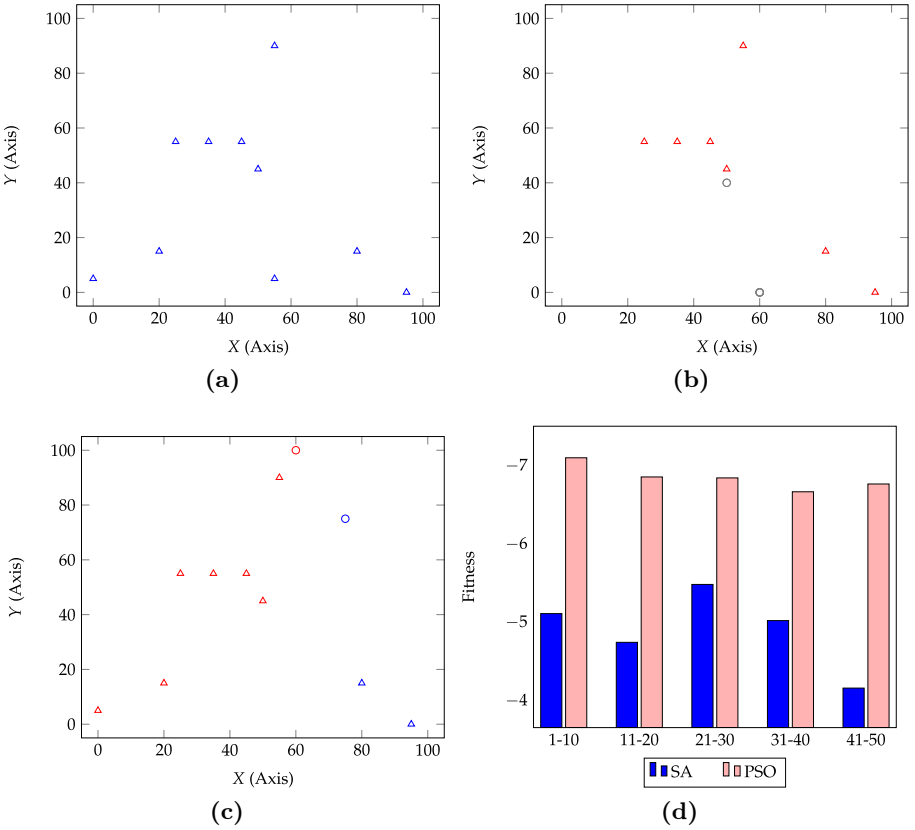


Fig. 4. Illustrations of: 4a Initial target positions; 4b Random initial solution; 4c Final solution found by PSO; 4d PSO vs SA

Finally, note that compared to *Y. Morsly et. al.* approach in [6] and *Yi-Chun Xu et al.* in [12], we consider a coverage quality parameter and deal with targets instead of area coverage. Here we must determine camera's positions and orientation, while coverage quality can be specified by the user. Even if we assume mobile targets, the proposed method can easily re-adapt the solution to the new scenario; this is guaranteed by the camera's movement (displacement and rotation).

7 Conclusion

We have proposed a mathematical mixed-integer model for the deployment of camera-based wireless sensor networks. Solving such models with exact is only feasible with small problem instances, beyond which, the use of approximative method such as meta-heuristics is required. In literature, only a few papers treat the target-based coverage model along with a coverage quality parameter. Hence, we have introduced an additional problem parameter to account for coverage quality. We have adapted a PSO algorithm to minimize the number of active cameras used to cover targets with certain quality. The proposed method gives easily a significant improvement of the initial random solution in few steps. Compared to SA, our PSO returns better solutions. We envisage later to calculate exact solutions by a ILP solver and then compare the results with ours. Moreover, we are working to introduce connectivity and energy-efficiency criteria to our problem. Finally, we are considering to use parallel versions of PSO and evaluate their performance over multiple platforms.

References

1. Ai, J., Abouzeid, A.A.: Coverage by directional sensors in randomly deployed wireless sensor networks. Journal of Combinatorial Optimization 11(1), 21–41 (2006)
2. Aziz, N.A.B.A., Mohemmed, A.W., Alias, M.Y.: A wireless sensor network coverage optimization algorithm based on particle swarm optimization and voronoi diagram. In: International Conference on Networking, Sensing and Control, ICNSC 2009, pp. 602–607. IEEE (2009)
3. Gorse, D.: Binary particle swarm optimisation with improved scaling behaviour. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2013)
4. Kannan, A.A., Mao, G., Vucetic, B.: Simulated annealing based wireless sensor network localization with flip ambiguity mitigation. In: IEEE 63rd Vehicular Technology Conference, VTC 2006-Spring, vol. 2, pp. 1022–1026. IEEE (2006)
5. Megerian, S., Koushanfar, F., Qu, G., Veltri, G., Potkonjak, M.: Exposure in wireless sensor networks: theory and practical solutions. Wireless Networks 8(5), 443–454 (2002)
6. Morsly, Y., Aouf, N., Djouadi, M.S., Richardson, M.: Particle swarm optimization inspired probability algorithm for optimal camera network placement. IEEE Sensors Journal 12(5), 1402–1412 (2012)
7. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intelligence 1(1), 33–57 (2007)

8. Trucco, E., Umasuthan, M., Wallace, A.M., Roberto, V.: Model-based planning of optimal sensor placements for inspection. IEEE Transactions on Robotics and Automation 13(2), 182–194 (1997)
9. Van Laarhoven, P.J., Aarts, E.H.: Simulated annealing. Springer (1987)
10. Veltri, G., Huang, Q., Qu, G., Potkonjak, M.: Minimal and maximal exposure path algorithms for wireless embedded sensor networks. In: Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, pp. 40–50. ACM (2003)
11. Wang, B.: Coverage problems in sensor networks: A survey. ACM Computing Surveys (CSUR) 43(4), 32 (2011)
12. Xu, Y.C., Lei, B., Hendriks, E.A.: Camera network coverage improving by particle swarm optimization. Journal on Image and Video Processing 3 (2011)

A version of LEACH Adapted to the Lognormal Shadowing Model

Chifaa Tabet Hellel[1], Mohamed Lehsaini [1(✉)], and Hervé Guyennet [2]

[1] STIC Laboratory, University of Tlemcen, Tlemcen, Algeria
[2] FEMTO-ST/DISC UFR ST, University of Franche-Comte, Besançon, France
tabetchifaa@gmail.com, m_lehsaini@mail.univ-tlemcen.dz,
herve.guyennet@femto-st.fr

Abstract. The most protocols designed for wireless sensor networks (WSNs) have been developed for an ideal environment represented by unit disc graph model (UDG) in which the data is considered as successfully received if the communicating nodes are within the transmission range of each other. However, these protocols do not take into account the fluctuations of radio signal that can happen in realistic environment. This paper aims to adapt LEACH protocol for realistic environment since LEACH is considered as the best cluster-based routing protocol in terms of energy consumption for WSNs. We have carried out an evaluation of LEACH based on two models; lognormal shadowing model (LNS) in which the probability of reception without error is calculated according to the Euclidian distance separating the communicating nodes and probabilistic model in which the probability of reception is generated randomly. In both models, if the probability of successful reception is lower than a predefined threshold, a multi-hop communication is incorporated for forwarding data between cluster-heads (CHs) towards the base station instead of direct communication as in original version of LEACH. The main aims of this contribution are minimizing energy consumption and guaranteeing reliable data delivery to the base station. The simulation results show that our proposed algorithm outperforms the original LEACH for both models in terms of energy consumption and ratio of successful received packets.

Keywords: LEACH · Lognormal shadowing model · Multi-hop scheme · Probabilistic model · Unit Disc Graph model · WSNs

1 Introduction

WSNs are composed of hundreds and thousands of small devices called "sensor nodes" distributed over a monitoring area for sensing data and sending it to a remote base station directly or via a multi-hop communication scheme depending on the application designed [1]. This novel technology has allowed the appearance of many applications such as; military, security, medical, environment monitoring, etc, due to the low cost of sensor nodes. Moreover, with this technology our way of life has been revolutionized since it allowed us to interact with the surrounded environment.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 465–475, 2015.
DOI: 10.1007/978-3-319-19578-0_38

Routing process is a fundamental operation in wireless sensor networks. It consists in establishing path to transmit a message from a source node to a remote base station according to the main routing schemes: hierarchical, location-based, data-centric and QoS-aware [2]. However, cluster-based routing in wireless sensor network is considered as the perfect solutions for minimizing energy consumption [3,4]. In this scheme, the network is divided into clusters wherein each cluster contains a number of members which sense data from its environment and send it to its corresponding cluster-head (CH). The latter is responsible for gathering data received from its members. If the distance between the source node and the destination node will increase, the energy consumption also increases, thereby a cluster-based routing scheme is recommended. Among the protocols proposed, LEACH (Low Energy Adaptive Clustering Hierarchy) [5] is considered as the best cluster-based protocol for saving energy. Nevertheless, the performance of these protocols may degrade in non-ideal environments.

In this paper, we used the lognormal shadowing model [6] and the probabilistic model to simulate a non-ideal environment, and we evaluated the performance of LEACH with these both models. Then, we proposed an improved version of LEACH to overcome the limitations of the original version. The proposed version involves a CH-to-CH routing scheme to guarantee reliable delivery. This routing scheme is used if the probability of reception of packets without error between cluster-heads and the base station is lower than a predefined threshold. Moreover, this scheme also permits to minimize energy consumption.

The rest of paper is organized as follow; in section 2, we give an overview on LEACH protocol and discuss some works that improve LEACH related to our requirements. Section 3 presents our improved version of LEACH to be adapted in realistic environment, and in section 4, we illustrate performance of LEACH and the proposed contribution in non-ideal environment. Finally, in section 5, we conclude our paper.

2 Related Work

Since LEACH protocol is considered among the best cluster-based routing protocols in terms of energy efficiency, a lot of researches have been enhancing this protocol to reduce its limitations. In the following, we present briefly LEACH protocol and some variants of it.

LEACH [5] is a cluster-based routing protocol that aims to minimize energy consumption and thereby increasing network lifetime. In LEACH, the network is divided into clusters and each cluster is headed by a cluster-head which is elected by itself by generating a random number between 0 and 1. If the number generated is lower than a predefined threshold, the concerned node becomes a CH for the current round. The threshold is computed by each sensor node according to the equation (1).

$$T(i) = \begin{cases} \frac{p}{1-p*\left(r \bmod \frac{1}{p}\right)} & if \ \ i \in G \\ 0 & otherwise \end{cases} \quad (1)$$

where p is the percentage of cluster-heads, r is the current round, G is the set of nodes that have not been selected as cluster-heads in the last (1/p) rounds.

LEACH is performed in two phases: setup phase and steady phase. In setup phase, each CH broadcasts an advertisement message to construct its cluster, and each non-CH that receives this message joins the adequate cluster based on the RSSI (Received Signal Strength Indication) of the message received. Once the clusters are formed, a TDMA (Time Division Multiple Access) schedules are assigned to member nodes in each cluster. In steady phase, each member transmits its sensed data to its corresponding CH in its scheduled time-slot, and then the CH aggregates all data received from its members and sends it to the remote base station directly. To avoid interference between cluster-heads, each CH chooses a CDMA (Coding Division Multiple Access) code that is different from other clusters to communicate with the base station.

In [7], the authors have proposed a multihop routing scheme with lower energy adaptive clustering hierarchy called MR-LEACH. In this scheme, the CHs are responsible to aggregate data sensed by their members and act as relay nodes for remote CHs from the base station. MR-LEACH increases network lifetime since it uses a multihop routing scheme. This protocol is performed in three phases: cluster formation at lowest level, cluster discovery at different levels and scheduling. At the beginning of each round, cluster formation phase is lunched to construct a table in which each node maintains the information about its neighbors (node identifier, residual energy and node status) by using a 'Hello' message. Then, CHs broadcast a HEAD-MSG message in its vicinity and each non-CH chooses its respective CH among those in its neighborhood based on the strength of RSSI. In cluster discovery, the base station broadcasts its identifier (ID). Each CH that receives this broadcasted message records the identifier of the base station and replies by a beacon signal with its ID. CHs that are closest to the base station are in level one i.e. they could reach the base station by single hop. Then, BS broadcasts again a control message, the CHs that are at level two reply to this message forwarded by CHs of level one and the BS would record cluster-head's ID and its level. Similarly, this process is repeated until no new CH is discovered. After that, the BS will form a cluster of CHs. In scheduling phase; after cluster formation, a TDMA scheduling is used for communication between CHs at different levels.

In [8], the authors proposed MH-LEACH which uses a new scheme for multihop communication to minimize energy consumption. MH-LEACH is carried out in two phases to establish paths towards the base station. In the first phase, the cluster-heads are selected as in the original version of LEACH protocol, and then each CH broadcasts an announcement message within its vicinity. Each non-CH that receives this message chooses the closest one based on the RSSI of the message received. Moreover, the base station performs the same process. In the second phase; each CH sends its initial route to reach the base station and the latter send back the route to the CH to confirm that there is a route between the considered CH and the base station. Therefore, a routing table is created by each CH that contains a list of available routes to the base station and the shortest one is used.

In [9], the authors proposed an improved version of LEACH in which a multihop scheme is used and the election of the cluster-heads is done according to its residual

energy. In this protocol, a multihop routing scheme for intra-cluster communication and a chain structure routing scheme for inter-cluster communication. The proposed protocol is performed in three phases: cluster formation, data delivery and update of clusters. At the beginning of the first phase, each node generates a random number between 0 and 1 and compares it with a predefined threshold to which the energy factor is added as illustrated by the equation (2).

$$T(n) = \begin{cases} \frac{p}{1-p*\left(r \bmod \frac{1}{p}\right)} \cdot \frac{E_n}{E_{Average}} & if \ n \in G \\ 0 & otherwise \end{cases} \quad (2)$$

where E_n is the residual energy and $E_{Average}$ is the average residual energy of all nodes.

Each CH sets H to 0, where H is the number of hops from CH, and broadcasts a message containing (H=0, ID) in its vicinity. Each non-CH that receives this message joins the cluster to which it belongs the transmitter CH and sets its own CH to the CH of the message received and H to H+1, and PID with the ID of its parent. Then, this node also broadcasts a message containing (H_j,ID). If a node receives more than one message it compares its own H with H of the sending nodes and it updates its H with the lowest one among those received. For intra-cluster communication, each node sends its packet to its parent and this latter sends it to its parent until reach the CH, and in inter-cluster communication, CH transmits data in chain structure. The concerned CH chooses the closest CH from it as the next hop and this process is repeated until reach the base station. In update of clusters, at the end of a round the remaining energy of CH may not be sufficient for the next round so the CH must be replaced by the node with the greatest residual energy.

In [10], an enhanced version of LEACH is proposed whose aims are saving energy by using a CH-to-CH multihop scheme and creating a backup path between cluster-heads to achieve fault-tolerance in the presence of failures. At the beginning, the base station broadcasts a HELLO message and each CH that receives this message calculates the RSSI. If RSSI is higher than a predefined threshold, this CH is closest to the base station and if not the CH is away from it and therefore it needs a relay node to reach the base station. The relay node is one of the CHs that are closest to the base station and it is selected based on the RSSI of a message exchanged between CHs and a variable called C_{red} which is a random number comprised between 0 and 1. If each of these parameters is higher than a threshold, this CH is considered as perfect relay node. However, the failure of one CH on the multihop path can affect the entire path and then the information cannot reach the base station, in this case a backup path is incorporated to this multihop path to ensure fault-tolerance and reliable delivery.

3 Contribution

Before presenting our contribution, we give a brief description of the lognormal model and probabilistic model. Then, we evaluate the performance of LEACH protocol with the both models to point out its weaknesses over an ideal environment.

3.1 Lognormal Shadowing Model

The lognormal shadowing model [6] is considered as a realistic model. It takes into account the fluctuations of radio signal caused by several factors such as noise, the presence of obstacles, weather conditions, etc... to evaluate the link quality between communicating nodes. The link quality is used to determine the probability of successful reception between communicating nodes in order to know if the message is received or it is corrupted by the destination node. Since this probability implied several factors, it may be difficult to obtain an accurate evaluation for all these factors which are themselves prone to errors. Therefore, we assume that signal strength gradually decreases according to the distance; thereby the probability of reception without errors can be computed according to the distance separating two nodes. We used the fluctuation of the signal model described in [11] as presented by the equation (3).

$$F(x) = \begin{cases} 1 - \frac{\left(\frac{x}{R_c}\right)^{2\alpha}}{2} & if \;\; 0 < x \leq R_c \\ \frac{\left(\frac{2R_c - x}{R_c}\right)^{2\alpha}}{2} & if \;\; R_c < x \leq 2R_c \\ 0 & otherwise \end{cases} \tag{3}$$

where α is the attenuation factor that depends on the environment and x is the distance separating the two communicating nodes. R_c is the transmission range and if the distance between two nodes is equal to R_c, the probability of successful reception is 0.5.

3.2 Probabilistic Model

In this model, the probability of reception without errors is generated randomly between each two communicating nodes. This model is used to illustrate the link quality. Moreover, the probability of successful reception is independent of the distance separating the communicating nodes but it depends on the factors which exist in the environment such as the presence or the absence of obstacles. Fig. 1 shows that node A can communicate with the node B but it cannot communicate with the node C although the distance that separates it with the node C is lower than that of the node B.

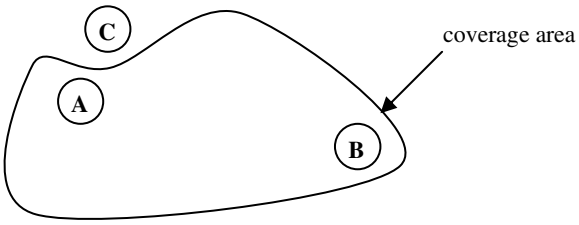


Fig. 1. Communication in probabilistic model

3.3 Proposed Scheme

In WSNs, LEACH is considered one of the best protocols in terms of energy efficiency. Several proposed protocols compare its effectiveness to LEACH and a lot of improved versions of LEACH have been proposed to reduce its limitations but they cannot guarantee its functionalities in a realistic environment. Our proposed algorithm aims to minimize energy consumption and ensures reliable delivery to the base station in a realistic environment based on lognormal shadowing model and probabilistic model.

We have proposed a multihop scheme instead of direct communication scheme between clusterheads and the base station to overcome the limitations of LEACH, such as when a CH aggregates data received from its members it computes the probability of reception without error of its packet to the base station. If this probability is higher than a predefined threshold, the packet is received correctly via direct communication by the base station and if not it means that the communication link is unreliable and in this case a multihop communication scheme will be incorporated to ensure the successful reception of packets by the base station. The proposed routing schemes are performed according to the following algorithms:

Algorithm 1: Routing scheme based on LNS model

- CH_S: Set of clusterheads
- BS: Base station whose coordinates (x_b, y_b)
- d: Euclidean distance between CH(x,y) and BS
- CH_D: Set of clusterheads that can send data directly to the BS
- CH_R: Set of clusterheads that use relay nodes to reach the BS

Begin

$CH_D = \emptyset$

$CH_R = \emptyset$

For ($CH \in CH_S$) **do**

- CH(x,y) calculates the Euclidean distance that separates it from the BS

$$d = \sqrt{(x - x_b)^2 + (y - y_b)^2}$$

- CH computes the probability of reception without error

$$Pr(d) = 1 - \frac{\left(\frac{d}{R_c}\right)^{2\alpha}}{2}$$

if (Pr(d) > Threshold) **then**

$CH_D = CH_D \cup \{CH\}$

else

$CH_R = CH_R \cup \{CH\}$

end if

end For

- Let CH a clusterhead

if (CH ∈ CH_D) **then**

- CH sends directly aggregated data to BS

else

- CH selects a (CH_r ∈ CH_D) as relay node with minimum distance to BS

Min = ∞

For (CH_i ∈ CH_D) **do**

- Computes the distance between CH(x,y) and $CH_i(x_i, y_i)$

$$dd = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

if dd < Min **then**

Min = dd

CH_r = CH_i

end if

end for

- CHr is selected as relay node by CH

end if

End

Algorithm 2: Routing scheme based on probabilistic model

- CH_S: Set of clusterheads
- BS: Base station
- CH_D: Set of clusterheads that can send data directly to the BS
- CH_R: Set of clusterheads that use relay nodes to reach the BS

Begin

CH_D = ∅

CH_R = ∅

For (CH ∈ CH_S) **do**

- CH generates a random number (rd_number) comprise between 0 and 1

if (rd_number ≥ 0.5) **then**

$CH_D = CH_D \cup \{CH\}$

else

$CH_R = CH_R \cup \{CH\}$

end if

end For

- Let CH a clusterhead

if (CH ∈ CH_D) **then**

- CH sends directly aggregated data to BS

else

- CH selects a (CH_r ∈ CH_D) as relay node such as CH and CH_r have a maximum probability.

pr = 0

For (CH_i ∈ CH_D) **do**

- CH generates a random number (rd_number) between CH and CH_i

if (rd_number > pr) **then**

pr = rd_number

CH_r = CH_i

end if

end for

- CH_r is selected as relay node by CH

end if

End

4 Simulation Results

Several simulations have been carried out to illustrate the performance of our contribution using TOSSIM simulator [12], and compared them with the original version of LEACH in terms of energy consumption and the ratio of successful received packets at the base station. For that, we used a network that contains respectively 20, 40, 60, 80 and 100 stationary nodes, which are randomly deployed on a 100m x 100m square area and the initial energy of each node is equal to 2 joules. The simulations were performed in 600 seconds, and we used a threshold p= 0.7 for probability of reception without error in LNS model. We used this threshold to avoid on the one hand the ideal model whose threshold is 0.5 and the other to avoid a highly disturbed environment. Moreover, for probabilistic model, we used a threshold of p=0.5 i.e. a clusterhead generates a random number comprise between 0 and 1 and if this number is higher than 0.5 we assume that this clusterhead can communicate directly with the base station. Table I summarize simulation parameters.

Table 1. Simulation Parameters

Parameter	Value
Deployment Area	100m x 100m
Simulation Time	600 sec
Number of nodes	20, 40, 60, 80, 100
Packet size	29 bytes
Initial node energy	2 Joules
Threshold for LNS model	p = 0.7
Threshold for probabilistic model	p = 0.5

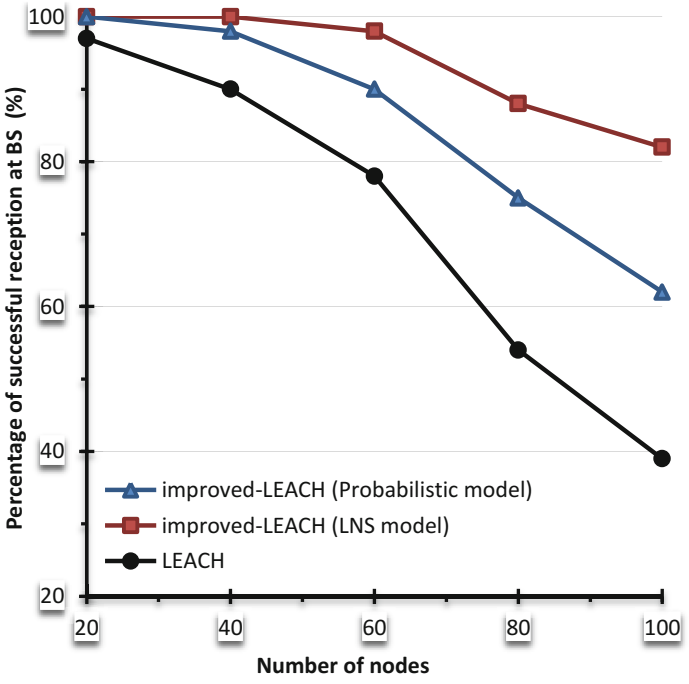


Fig. 2. Evaluation of ratio of successful received packets at BS with p=0.7

Fig. 2 shows that the ratio of successful packets received at base station with a probability of p=0.7 in improved LEACH is higher than in original LEACH and also the ratio is higher with probabilistic model compared with LEACH. In improved LEACH the unreliability of links between a clusterhead and the base station can be treated by a multihop communication by against, in original LEACH the packet will be lost due to the unreliable links.

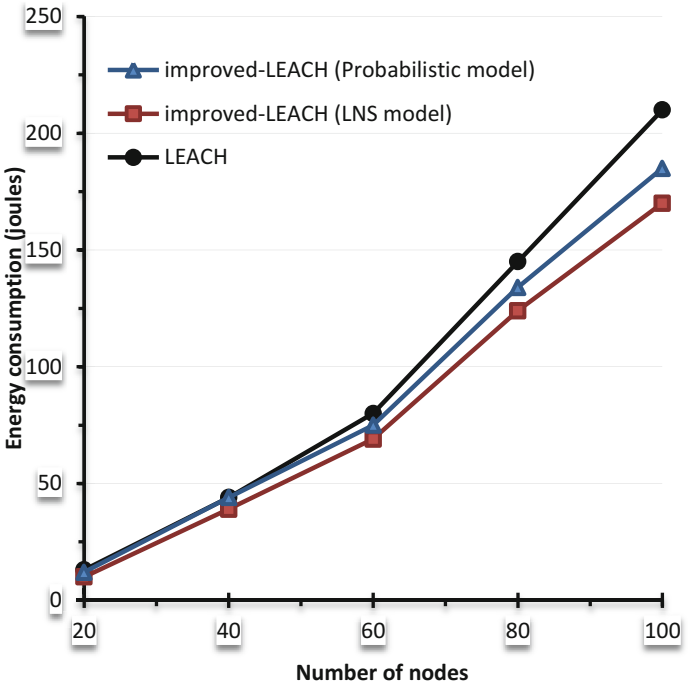


Fig. 3. Energy consumption in LEACH and Improved-LEACH

Fig.3 illustrates that energy consumption in improved LEACH based on LNS model or Probabilistic model is lower than in LEACH because in the improved version, the multihop transmission scheme minimizes energy consumption but the direct communication in LEACH consumes more energy.

5 Conclusion

In this paper, we have evaluated LEACH protocol in a realistic environment represented by lognormal shadowing model and a probabilistic model. However, results obtained illustrated that the performance of LEACH degrades in this kind of environment. Therefore, we have proposed an improved version of LEACH to overcome its weaknesses with realistic environment. The proposed scheme aims to find an optimal multihop path over links error which is modeled by LNS model and probabilistic model.

The simulation results showed that our contribution outperforms the original version of LEACH in terms of energy consumption and ratio of successful packets received at the base station. Moreover, our contribution deals with fault tolerance in LEACH, ensures reliable delivery and minimizes energy consumption.

References

1. Chanak, P., Banerjee, I.: Energy efficient fault-tolerant multipath routing scheme for wireless sensor networks. The Journal of China Universities of Posts and Telecommunications 20(6), 42–61 (2013)
2. Akkaya, K., Younis, M.A.: Survey on routing protocols for wireless sensor networks. Ad Hoc Networks 3(3), 325–349 (2005)
3. Tyagi, S., Kumar, N.: A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks. Journal of Network and Computer Applications 36(2), 623–645 (2013)
4. Vlajic, N., Xia, D.: Wireless sensor networks: to cluster or not to cluster? In: International Symposium on World of Wireless, Mobile and Multimedia Networks, pp. 260–268 (2006)
5. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd IEEE Annual Hawaii International Conference on System Sciences, Hawaii, pp. 1–10 (2000)
6. Rappaport, T.S.: Wireless Communications Principles and Practice, 2nd edn. Prentice Hall Release (2001)
7. Farooq, M.O., Dogar, A.B., Shah, G.A.: MR-LEACH: Multi-hop Routing with Low Energy Adaptive Clustering Hierarchy. In: Proceeding of Fourth IEEE International Conference on Sensor Technologies and Applications (IEEE), Venice, Italy, pp. 262–268 (2010)
8. Neto, J., Antoniel Rego, A., Andr-Cardoso, A., Jnior, J.: MH-LEACH: A Distributed Algorithm for Multi-Hop Communication in Wireless Sensor Networks. In: Proceeding of The Thirteenth International Conference on Networks (ICN), Nice, France, pp. 55–61 (2014)
9. Yang, H., Xu, J., Wang, R., Qian, L.: Energy-Efficient Multi-hop Routing Algorithm Based on LEACH. In: Wang, R., Xiao, F. (eds.) CWSN 2012. CCIS, vol. 334, pp. 578–587. Springer, Heidelberg (2013)
10. Tabet Hellel, C., Lehsaini, M., Guyennet, H.: An Enhanced Fault-tolerant Version of LEACH for Wireless Sensor Networks. International Journal of Advancements in Computing Technology(IJACT) 6(6), 50–57 (2014)
11. Kurvilla, J., Nayak, A., Stojmenoviç, I.: Hop count optimal position based packet routing algorithms for ad hoc wireless networks with a realistic physical layer. IEEE Journal on Selected Areas in Communications 23(6), 1267–1275 (2005)
12. Levis, P., Lee, N., Welsh, M., Culler, D.: TOSSIM: accurate and scalable simulation of entire TinyOS applications. In: The First ACM International conference on Embedded networked sensor systems (SenSys 2003), New York, USA, pp. 126–137 (2003)

Security and Network Technologies: Energy and Synchronisation

High Velocity Aware Clocks Synchronization Approach in Vehicular Ad Hoc Networks

Khedidja Medani(✉), Makhlouf Aliouat, and Zibouda Aliouat

Faculty of sciences, Ferhat Abbas University Sétif 1 Algeria
Khadidja-medani@hotmail.fr,
{aliouat_m,aliouat_zi}@yahoo.fr

Abstract. Clock synchronization plays an important role in communications organization between applications in Vehicular Ad hoc NETworks (VANETs) requiring a strong need for coordination. Having a global time reference or knowing the value of a physical clock (indeed with an acceptable approximation) of cooperative process involved in the provision of a service by distributed applications, takes on a fundamental importance in decentralized systems, particularly in VANETs. The intrinsic and constraining features of VANETs, especially the high mobility of vehicles make the clock synchronization mechanisms more complex and require a concise and a specific adequacy. The aim of the work reported in this paper is to propose a new protocol for clocks synchronization for VANETs, sufficiently robust, with a good precision, and convenient to the main constraint such high nodes mobility. Our proposed protocol, named Time Table Diffusion (TTD), was simulated using a combination of two simulators: VanetMobiSim and NS2 to evaluate its performance in terms of convergence time and number of messages generated. The obtained results were conclusive.

Keywords: VANETs · Clocks synchronization · Intelligent Transportation System · Worthwhile Road Traffic · Time Table Diffusion · TTD

1 Introduction

Over the last decade, the use of wireless ad hoc network in transportation domain has drawn particular researchers' attention in order to promote them to a satisfactory rank regarding to the numerous advantages they may provide. So, communications between vehicles (IVC: Inter Vehicular Communications) have becoming one of the most active researching area. This applicative aspect has given a new communication paradigm that ensures to the classical transportation systems more efficiency, security, conviviality, and performances. So this gives rise to the so-called intelligent transportation systems (ITS). Although vehicular ad hoc networks (VANETs) as well as Wireless Sensor Networks (WSNs) are derived from the same source namely Mobile Ad hoc NETwoks (MANETs), the satisfying results obtained from researches and works done in these fields cannot be directly applied in the context of VANETs, because the specificities of the latter are more stringent in one side and plentiful in the

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 479–490, 2015.
DOI: 10.1007/978-3-319-19578-0_39

other. For example, the velocity of nodes in VANETs may reaches extreme values while energy is abundant and does not represent any constraint. So, the high mobility environments related to road infrastructure imposes new constraints like radio obstacles, the effects of multipath and fading.

Various common services such as communication, coordination, security, and time distribution channel access method for time slot (TDMA: Time Division Multiple Access) depend strongly on the existence of synchronized clocks of different nodes (vehicles) of a considered VANET network. Thus, clock synchronization requires the availability of a common time reference for all vehicles, and since these clocks drifted naturally, it is crucial to realize synchronization with an appropriate period and accuracy.

In contrast to other dynamic networks, high mobility of VANETs imposes new requirements in terms of immediate reactivity and high dynamic connectivity. Consequently, the clock synchronization methods used in Ad Hoc networks (MANETs and WSNs) are not suitable, it is therefore important to adapt them specifically to the context of VANET or proposing new well suited. Few works devoted to the problematic of clocks synchronization in VANETs were reported in the literature, such as: RBS [1] CTS [2], TTT [3] and HCS [4].

The aim of the work reported in this paper is to propose a new protocol, for synchronizing node's clocks in VANETs, independently of the network topology, based on a decentralized approach, and requiring no use of a Global Positioning System (GPS) component or an existing infrastructure. The proposed protocol should be able of providing debrided synchronization where each node moves freely with the time of its local clock, but stores the needed data to synchronize other nodes. It should also provide a good precision (of the order of micro seconds), robustness against failure of nodes, and a low cost in terms of convergence time and number of messages generated.

This paper is organized as follows: After an introduction of the problematic in Section 1, Section 2 presents previous work related to clock synchronization in VANETs. Section 3 is devoted to the presentation of our proposition (TTD: Time Table Diffusion), while Section 4 is dedicated to the exhibition of simulation results of TTD. We conclude our work with a conclusion and future perspectives.

2 Related Work

Several protocols for clock synchronization in VANETs have been proposed. These protocols are classified into two approaches (Fig. 1):

2.1 Centralized Approach

Among the proposed algorithms in centralized approach include GNSS: Global Synchronization for Satellite Navigation System [5] and Synchronization in ad hoc networks based on UTRA TDD [6]. These algorithms have the advantage of implementation simplicity, but however require the use of a GPS component, which may raise the problem of transmission signals power which may interfere with communications in progress within nodes.

2.2 Decentralized Approach

Decentralized synchronization algorithms are sufficient for inter vehicular communications and better than the centralized ones in terms of fault tolerance. These algorithms are classified into three categories according to the time information exchange mode between vehicles [7] and are as follows:

- Burst position measurement: Each node programs the periodic transmission of a pulse and corrects its own local after receiving the new burst.
- Continuous correlation of timing signals: Each node continuously transmits a signal sequence and calculates the phase offset using the received sequence. Examples of synchronization protocols based on this method are presented in [8] [9].
- Clock-sampling methods: Each node reads its clock time and transmits it explicitly to other neighboring nodes. At each reception, the offsets are calculated as the difference between the local time and the time clocks of neighboring nodes. This method is superior to the other two methods in terms of simplicity, because it directly exchanges time information, without regard to phase. Among the protocols based on this method include those described in [1] [2] [3].

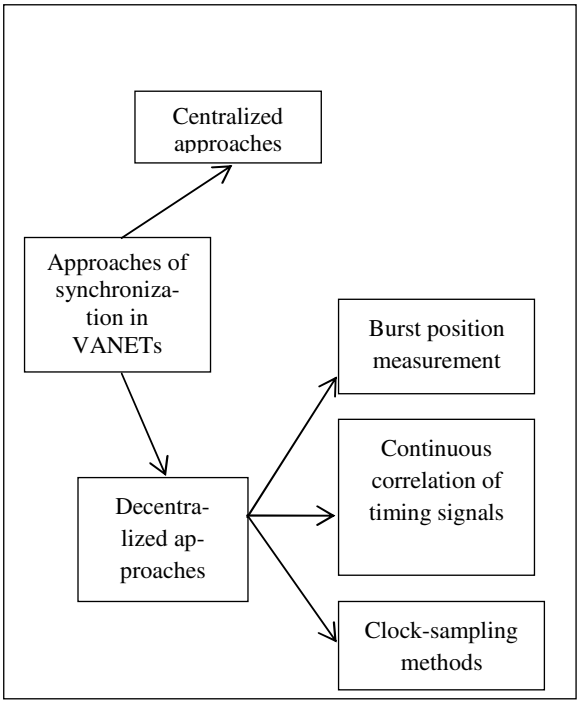


Fig. 1. Clock synchronization protocols classification in VANETs

3 Clock Synchronization with Time Table Diffusion Protocol

Our proposition named Time Table Diffusion (TTD) exploits the idea of transferring a time table implemented by the TTT protocol [3] for clock synchronization in mobile sensor networks. The basic idea is to choose a transporter node (T) to transfer a time table containing the offsets related to different nodes. These offsets are calculated by the offset delay estimation method [10]. Transferring time table by the transporter node makes nodes able to calculate their relative clock offsets with the nodes in the time table without even having any message exchanges. Thus, this will offers a great advantage since it contributes to avoid network congestion.

TTD provides synchronization in vehicular environments independent of the network topology in which each node has a unique identity in the network.

TTT protocol uses nodes mobility to transfer time table. The clock offset associated to each node will be kept in the memory of node in a time table, and upon communicating with a new node, the time table would be transferred to the other node. This process provides a long convergence time (in order of seconds) which make a conflict with the real time applications of VANETs (alert messages …).

To explain the functioning of TTD, synchronization steps are illustrated in Fig. 2.

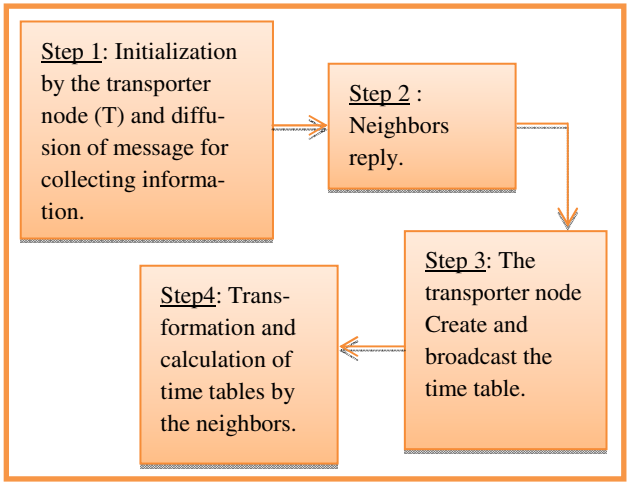


Fig. 2. Synchronization steps in TTD

The synchronization process begins with broadcasting of a message by the transporter node for collecting information. Neighboring nodes respond to transporter node to construct the time table (Time_table). Once the latter one is built, it should be broadcasted later by the transporter node. We describe these steps in the sequel:

3.1 Step 1

In this step, the transporter node broadcasts an advertisement message to initiate the synchronization process and to collect information needed to build the time table. The broadcasting message contains the identity of the transporter node T and t_0, the time-stamp indicating the sending instant of this message.

Each node begins this step with sending CTS (Clear to Send) messages. The first node sending its CTS becomes the transporter node in its neighborhood.

3.2 Step 2

A node i that receiving the advertisement message of the transporter node T marks it at the receiving instant of t_{1i}, and then sends a response message to the transporter node T. The response message contains the identity i of the node, and the timestamps t_0, t_{1i}, and t_{2i} where t_{2i} represent the instant of sending response message. One node i may join more than one transporter node at the same time.

3.3 Step 3

When T receives the reply from node i at the instant t_{3i}, using timestamps t_0, t_{1i}, t_{2i}, and t_{3i}, T can calculate the offset relative to node i (Δ_{iT}) according to the equation (1) below and saves the result in the time table where the index access is the identity of node i.

$$\Delta_{iT} = ((t_{1i} - t_0) - (t_{3i} - t_{2i}))/2. \quad (1)$$

Fig. 3 hereafter illustrates the messages exchange between the transporter node T and a node i:

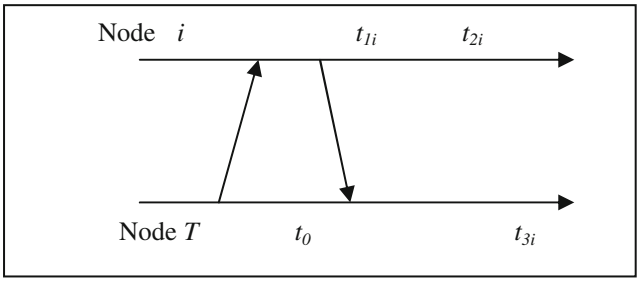


Fig. 3. Messages exchange between the transporter node *T* and a node *i*

After T has completed the construction of the time table, it has to broadcast it to all its neighbors' node allowing them to build their own time tables.

3.4 Step 4

When a node i receives the time table from a transporter node T, it can build its own table as follows:

Node i will search in the received table the corresponding value to its identity (Δ_{iT}), and stores the inverse of this value in its own table in the location corresponding to the identity of the node T (*time_table(T)*=-Δ_{iT}). The principle is the following:

$$\Delta_{iT} = C_i - C_T. \tag{2}$$

Multiplying both sides of (2) by (-1), we obtain:

$$\Delta_{Ti} = C_T - C_i. \tag{3}$$

Where, C_T is the clock value of node T at the instant t, and C_i is the clock value of node i in the same instant t.

To synchronize itself with the rest of the nodes table, node i will add Δ_{Ti} value to all values in the table according to the following principle:

$$\Delta_{jT} = C_j - C_T. \tag{4}$$

$$\Delta_{Ti} = C_T - C_i. \tag{5}$$

By adding the two parts of (4) and (5) we obtain:

$$\Delta_{jT} + \Delta_{Ti} = (C_j - C_T) + (C_T - C_i) = \Delta_{ji}. \tag{6}$$

As shown in Fig. 4, depending on the transporter node (that depend on the random number generated by each node), we can find two neighbors not synchronized (node 2 and node 4 participate to the synchronization process under different transporter nodes, that make nodes 2 and 4 two neighbors not synchronized).

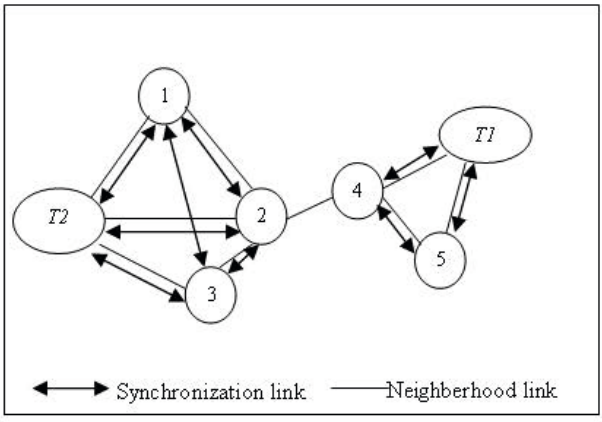


Fig. 4. Problem posed by the random time

A solution to this problem is that inspired from [8] which consist to larger the range of synchronization packet transmission to be equal double that of data packet transmission. In this way, a transporter node T ensures the synchronization of the nodes joining with all its neighbors (one hop) and in most cases, the synchronization on multi-hop paths.

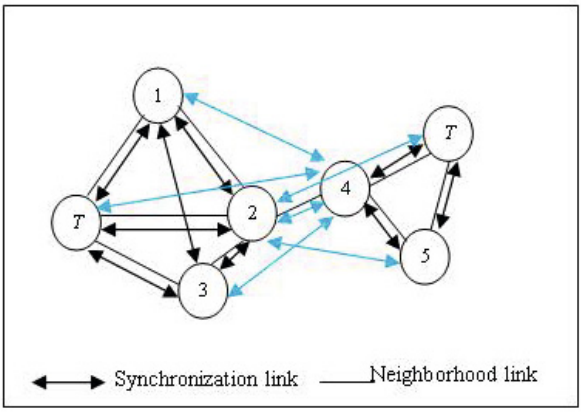


Fig. 5. Improved initial model of the synchronization by TTD

For mobility management, and since the clocks drifted naturally (the live duration of synchronization is an important evaluation criteria of synchronization's algorithms), it is crucial to achieve often synchronization process cycles as shown in Fig. 6.

Table 1. Exchange message type and their content

Message	***Number***	***Content***
The avertissement message sent by the transporter node (ADV_T).	n_t, where n_t is the number of transporter nodes in the current cycle.	Transporetr node identity and the timestamp t_0.
Neighbords reply (JOIN_RESPONSE).	$\sum_{0}^{n_t-1} N_i$	Neignbord i identity and timestamps t_0, t_{1i}, t_{2i}.
Time table (TIME_TABLE).	n_t	Time table built by the transporter node

The number of messages necessary to accomplish the synchronization is calculated as follows: Assuming there are Ni nodes within the synchronization scope of a transporter node Ti, where Ti ∈ T (where T is the set of transporter nodes in the current cycle). We can summarize the number and content of messages required for synchronization, as shown in Table I.

According to this table, the number of messages (nbMsg) necessary to accomplish the synchronization can be estimated as follows:

$$nbMsg = \sum_{i=0}^{n_t-1} (N_i+2) \quad (7)$$

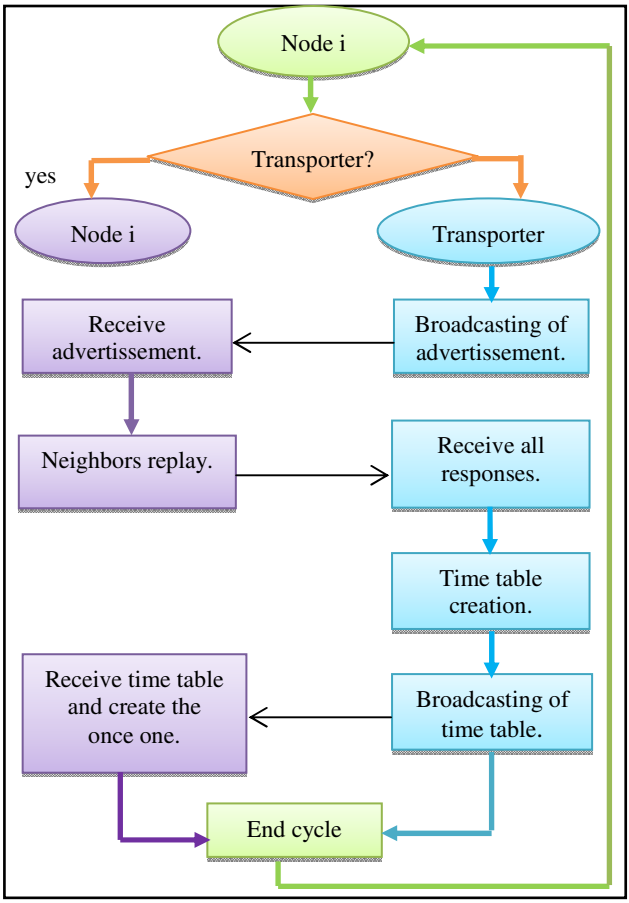


Fig. 6. Clock synchronization using TTD

4 Simulation Results

We simulated the proposed protocol using the combination of the simulator NS2 and the mobility generator VanetMobiSim. Clocks values used in simulation are randomly generated according to the law of GAUSS (0 average, $\delta = 10$ ppm) [11].

We tested a number of scenarios by changing essential parameters to evaluate the performances of our proposed protocol where nodes are initially placed in random positions and their movement direction follows the mobility model implemented by VanetMobiSim in Intelligent Driver Model with Lane Changing (IDM_LC: It regulates vehicle speed based on movements of neighboring vehicles (e.g., if a car in

front brakes, the succeeding vehicles also slows down). The implementation reflects restrictions of the spatial environment. Vehicles moving according to the IDM_LC model support smart intersection management: they slow down and stop at intersections, or act according to traffic lights, if present. The implementation reflects restrictions of the spatial environment. Also, vehicles are able to change lane and perform overtaking in presence of multi-lane roads).

Table 2. Simulation Parameters

Topologie (m2)	1000*1000
Nodes number	30/50/100/200/300
Speed (m/s)	7/10/15/20/25/30/35
Trafic light	6
Mobility model	Randomly according to IDM_LC with 2 obstacles every 100 m^2
Range data transmission (m)	250/500/1000
Simulation time (s)	1000

The metrics used to analyze the simulation results are the number of messages generated and the time of convergence (convergence time is the time required to accomplish the synchronization process).

Fig. 7 shows that the convergence time in TTD increases with increasing of nodes number, this is due to the large number of neighbors reply messages. In contrast, nodes speed has no influence on the convergence time because TTD solution uses broadcast (Fig.8). This property is an advantage for the proposed algorithm and makes it usable in different vehicles mobility environments (urban, suburban, and highway).

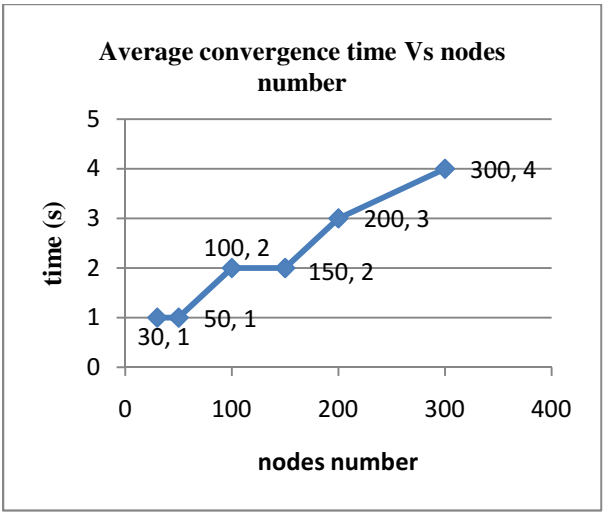


Fig. 7. Convergence time Vs nodes number in TTD

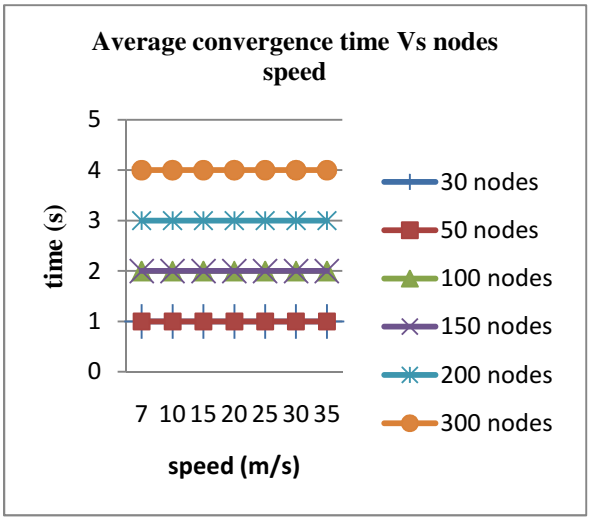


Fig. 8. Convergence time Vs nodes speed in TTD

However, convergence time in TTT [3] increases in urban environments characterized by a minimal speed compared to other vehicles mobility environments. This is because the TTT protocol uses node mobility as an essential factor for time table transfer. Thus, the convergence time shown by our protocol is less than that shown by the reference protocol TTT under the same conditions, as shown in Fig. 9.

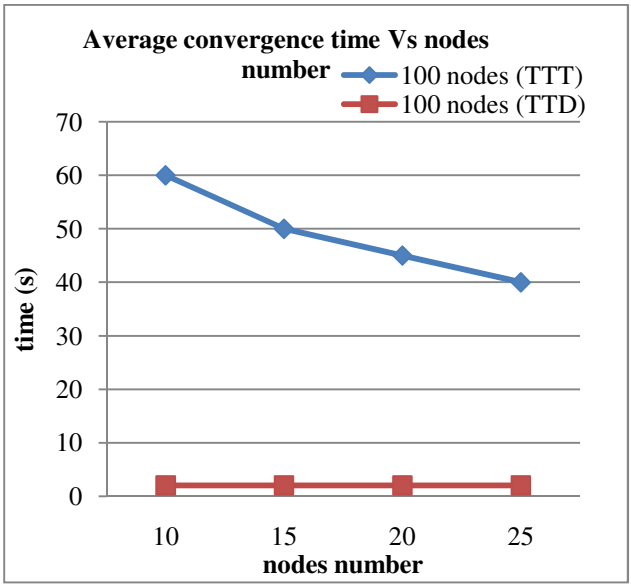


Fig. 9. Convergence time in TTD Vs Convergence time in TTT

The number of messages required to accomplish the synchronization process is not fixed and depends on two essentials factors; nodes number and transporter nodes number (that depends on the transmission range). On one hand, the number of messages increases with a large number of nodes (as shown in Fig. 10); logically this is due to the phase of neighborhood replays, the most consuming phase in the synchronization process in term of messages number.

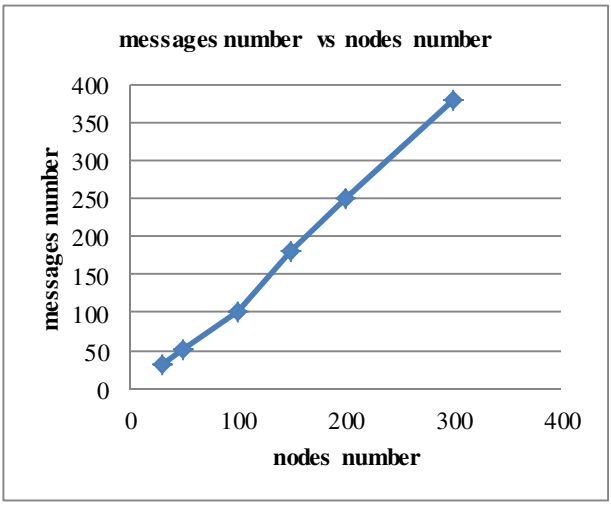


Fig. 10. Messages number Vs nodes number

On the other hand, depending on the transmission range, that affects the transporter nodes number, the number of messages increases with a the increasing of the number of transporter nodes. For example, in a topology 1000 * 1000 (m^2) with a number of nodes equal to 30 (low density network), we can achieve a data transmission range up to 1000 m, in this case, only one transporter node is sufficient to achieve the synchronization process, so we can reach a minimum number of messages that is equal to the nodes number in the network plus one.

5 Conclusion

Although clock synchronization in VANETs is a very important research area, few works have been reported so far in the specialized literature. In this paper, we propose a new efficient synchronization protocol taking into account the specific constraints imposed by VANET environments. The proposed solution called TTD (Time Table Diffusion) provides released clock synchronization in a vehicular environment independently of the network topology. TTD achieves synchronization with a good accuracy of the order of a microsecond, and in most cases, synchronization of multi-hop paths. The proposed solution is simulated with the combination of VanetMobiSim-NS2 to evaluate its performance in terms of number of messages generated and

convergence time. The simulation results showed that TTD provides a best convergence time compared to its homologues TTT (TTT provide best result than RBS in term of convergence time).

References

1. Elson, J., Girod, L., Estrin, D.: Fine-Grained Network Time Synchronization using Referece Broadcasts, vol. 36, pp. 147–163. ACM (2002)
2. Shizhun, W., Anjum, P., Maziar, N.: Converging Time Synchronization Algorithm for Highly Dynamic Vehicular Ad Hoc Networks (VANETs), vol. 6, pp. 443–448. IEEE (2010)
3. Reza, K., Lim, M., Sim, H., Tat Ewe, A.T., Wei, S.: Time Table Transfer Time Synchronization. In: Mobile Wireless Sensor Networks, vol. 5. PIERS Proceedings, Beijing (2009)
4. Sam, D., Cyril Raj, V.: A Time Synchronized Hybrid Vehicular Ad Hoc Network of Roadside Sensors and Vehicles for Safe Driving. Journal of Computer Science 11, 1617–1627 (2014)
5. Scopigno, R., Cozzetti, H.: GNSS synchronization in Vanets. In: 2009 3rd International Conference on IEEE New Technologies, Mobility and Security (NTMS), vol. 5(11), pp. 1–5 (2009)
6. Ebner, A., Rohling, H., Halfmann, R., Lott, M.: Synchronization in ad hoc networks based on UTRA TDD. In: The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 4 (2002)
7. Rentel, C.H.: Network Time Synchronization and Code-based Scheduling for Wireless Ad Hoc Networks. Carleton University, Ottawa (2006)
8. André, E., Hermann, R., Matthias, L., Rüdiger, H.: Decentralized Slot Synchronizatio. In: Highly Dynamic Ad Hoc Networks, vol. 2, pp. 494 – 498. IEEE (2002)
9. Nakagawa, E., Sourour, M.: Mutual Decentralized Synchronization for Intervehicule Communications, vol. 48(16). IEEE (1999)
10. Mills, D.L.: Internet Time Synchronisation: The Network Time Protocol, vol. 39. IEEE (1991)
11. Lombardi, M.A.: Frequency Measurement. The Measurement, Instrumentation and Sensors Handbook. CRC Press (1999)
12. André, E., Lars, W., Hermann, R.: Aspects of Decentralized Time Synchronization in Vehicular Ad hoc Networks. In: 1st International Workshop on Intelligent Transportation, Humburg (2004)

An Energy-Efficient Fault-Tolerant Scheduling Algorithm Based on Variable Data Fragmentation

Chafik Arar(✉), Mohamed Salah Khireddine, Abdelouahab Belazoui, and Randa Megulati

Department of Computer Science, University of Banta
BATNA 05000, Algeria
chafik.arar@gmail.com,
{mkhireddine,belazoui}@yahoo.fr,
randa_meguellati@hotmail.fr
http://www.univ-batna.dz

Abstract. In this article, we propose an approach to build fault-tolerant distributed real-time embedded systems. From a given system description and a given fault hypothesis, we generate automatically a fault tolerant distributed schedule that achieves low energy consumption and high reliability efficiency. Our scheduling algorithm is dedicated to multi-bus heterogeneous architectures with multiple processors linked by several shared buses, which take as input a given system description and a given fault hypothesis. It is based on active redundancy to mask a fixed number L of processor failures supported in the system, and passive redundancy based on variable data fragmentation to tolerate N buses failures. In order to maximize the systems reliability, the replicas of each operation are scheduled on different reliable processors and the size of each fragmented data depends on GSFR and the bus failure rates. Finally, we show with an example that our approach can maximize reliability and reduce energy consumption when using active redundancy.

Keywords: Energy consumption · Scheduling · Embedded systems · Real time systems · Reliability · Active redundancy · Multi-bus architecture · variable data fragmentation

1 Introduction

Nowadays, heterogeneous systems are being used in many sectors of human activity, such as transportation, robotics, and telecommunication. These systems are increasingly small and fast, but also more complex and critical, and thus more sensitive to faults. Due to catastrophic consequences (human, ecological, and/or financial disasters) that could result from a fault, these systems must be fault-tolerant. This is why fault tolerant techniques are necessary to make sure that the system continues to deliver a correct service in spite of faults Jalote [1], [2],

A fault can affect either the hardware or the software of the system; we chose to concentrate on hardware faults. More particularly, we consider processors

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 491–502, 2015.
DOI: 10.1007/978-3-319-19578-0_40

and communication faults [3], [4]. In the literature, we can identify several fault-buses tolerance approaches for distributed embedded real-time systems, which we classify into two categories: proactive or reactive schemes.

In the proactive scheme [5], [6], multiple redundant copies of a message are sent along distinct buses. In contrast, in the reactive scheme only one copy of the message, called primary, is sent; if it fails, another copy of the message, called backup, will be transmitted. In [7], an original off-line fault tolerant scheduling algorithm which uses the active replication of tasks and communications to tolerate a set of failure patterns is proposed; each failure pattern is a set of processor and/or communications media that can fail simultaneously, and each failure pattern corresponds to a reduced architecture. The proposed algorithm starts by building a basic schedule for each reduced architecture plus the nominal architecture, and then merges these basic schedules to obtain a distributed fault tolerant schedule. It has been implemented in [8].

In [9], a method of identifying bus faults based on a support vector machine is proposed. In [2], faults of buses are tolerated using a TDMA (Time Division Multiple Access) communication protocol and an active redundancy approach. In [10] authors propose a fine grained transparent recovery, where the property of transparency can be selectively applied to processes and messages. In [11] authors survey the problem of how to schedule tasks in such a way that deadlines continue to be met despite processor and communication media (permanent or transient) or software failure.

In this paper, we are interested in approaches based on scheduling algorithms that maximize reliability and reduce energy consumption [12], [13], [14] when using active redundancy to tolerate processors faults and passive redundancy based on variable data fragmentation to tolerate buses faults.

The remaining of this paper is structured as follows: In section 2, we give detailed description of our system models. In section 3, we present our solution and we give detailed description of our scheduling algorithm. Section 4 shows with an example how our approach can maximize reliability and reduce energy consumption when using active redundancy. We finally conclude this work in section 5.

2 System Description

Distributed real-time embedded systems are composed of two principal parts, which are the algorithm (software part) and the distributed architecture (hardware part). The specification of these systems involve describing the algorithm (algorithm model), the architecture (architecture model), and the execution characteristics of the algorithm onto the architecture (execution model).

The algorithm is modeled as a data-flow graph noted ALG. Each vertex of ALG is an operation (task) and each edge is a data-dependence. A data-dependence, noted by $\rightarrow$, corresponds to a data transfer between a producer operation and a consumer operation. $t_1 \rightarrow t_2$ means that t_1 is a predecessor of t_2 and t_2 is a successor of t_1. Operations with no predecessor (resp. no successor) are the input interfaces (resp. output).

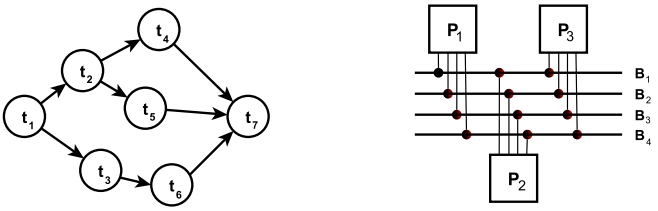


Fig. 1. ALG and ARC graphs

The architecture is modeled by a non-directed graph, noted ARC, where each node is a processor, and each edge is a bus. Classically, a processor is made of one computation unit, one local memory, and one or more communication units, each connected to one communication link. Communication units execute data transfers. We assume that the architecture is heterogeneous and fully connected. Figure 1 presents an example of ALG with seven operations t_1, t_2, t_3, t_4, t_5, t_6 and t_7 and ARC, with three processors P_1, P_2, P_3 and four buses B_1, B_2, B_3 and B_4.

Our real-time system is based on cyclic executive; this means that a fixed schedule of the operations of ALG is executed cyclically on ARC at a fixed rate. This schedule must satisfy one real-time constraint which is the length of the schedule. As we target heterogeneous architecture, we associate to each operation t_i a worst case execution time (WCET) on each processor P_j of ARC, noted $Exe(t_i, P_j)$. Also, we associate to each data dependency $data_i$ a worst case transmission time (WCTT) on each bus B_j of the architecture, noted $Exe(data_i, B_j)$.

We assume only processors and buses failures. We consider only transient bus faults, which persist for a short duration. We assume that at most L processors faults and N bus faults can arise in the system, and that the architecture includes more than L processors and N buses.

3 The Proposed Approach

In this section, we first discuss the basic principles used in our solution, based on scheduling algorithms. Then, we describe in details our scheduling algorithm. The aims of this algorithm are twofold, first, maximize the reliability of the system and minimize the length of the whole generated schedule in both presence and absence of failures; Secondly, reduce energy consumption. In our approach. we achieve high reliability, reducting consuption and fault tolerance in tow ways:

3.1 Active Redundancy with Changing Frequency

In order to tolerate up to L arbitrary processors faults, our solution is based on active redundancy approach. The advantage of the active redundancy of operations is that the obtained schedule is static; in particular, there is no need

for complex on-line re-scheduling of the operations that were executed on a processor when the latter fails; also, it can be proved that the schedule meets a required real-time constraint, both in the absence and in the presence of faults. In many embedded systems, this is mandatory. To tolerate up to L processors faults, each operation t of Alg is actively replicated on $L+1$ processors of Arc (see Figure 2). We assume that all values returned by the $L+1$ replicas of any operation t of Alg are identical.

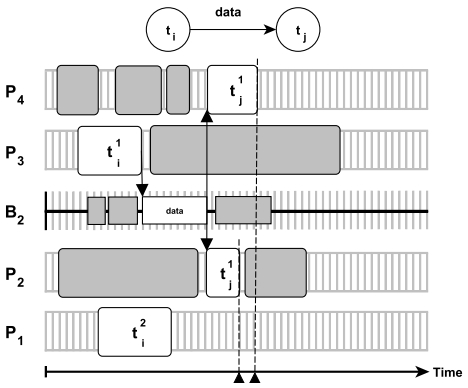


Fig. 2. Active redundancy

Voltage, Frequency and Energy Consumption: the maximum supply voltage is noted V_{max} and the corresponding highest operating frequency is noted F_{max}. For each operation, its WCET assumes that the processor operates at F_{max} and V_{max} (and similarly for the WCCT of the data-dependencies). Because the circuit delay is almost linearly related to 1/V, there is a linear relationship between the supply voltage V and the operating frequency F. In the sequel, we will assume that the operating frequencies are normalized, that is, $F_{max} = 1$ and any other frequency F is in the interval $[0, 1]$. Accordingly, the execution time of the operation or data-dependency M placed onto the hardware component C (be it a processor or a communication link) running at frequency F (taken as a scaling factor) is :

$$Exe(M, C, F) = \frac{Exe(M, C)}{F} \tag{1}$$

To calculate the power consumption, we follow the model presented in [15]. For a single operation placed onto a single processor, the power consumption P is :

$$P = P_s + h(P_{ind} + P_d) \tag{2}$$

Where P_s is the static power (power to maintain basic circuits and to keep the clock running), h is equal to 1 when the circuit is active and 0 when it is inactive, P_{ind} is the frequency independent active power (the power portion that

is independent of the voltage and the frequency; it becomes 0 when the system is put to sleep, but the cost of doing so is very expensive),

$$P_d = C_{ef} * V^2 * F \tag{3}$$

P_d is the frequency dependent active power (the processor dynamic power and any power that depends on the voltage or the frequency), C_{ef} is the switch capacitance, V is the supply voltage, and F is the operating frequency.

For processors, this model is widely accepted for average size applications, where C_{ef} can be assumed to be constant for the whole application. For a multiprocessor schedule S, we cannot apply directly the previous equation. Instead, we must compute the total energy *E(S)* consumed by S, and then divide by the schedule length *L(S)*:

$$P(S) = \frac{E(S)}{L(S)} \tag{4}$$

We compute *E(S)* by summing the contribution of each processor, depending on the voltage and frequency of each operation placed onto it. On the processor P_i, the energy consumed by each operation is the product of the active power $P^i_{ind} + P^i_d$ by its execution time.

In our approach, as *L+1* replicas of each operation are scheduled actively on *L+1* distinct processors, the energy consumed by the system is maximal. In order to reduce energy consumption, we propose to execute the *L+1* replicas of an operation with different frequencies F. As all the *L+1* replicas of an operation may have different end execution time (see Figure 2 for the replicas t^1_j and t^2_j), we choose to align the execution time of all the replica by changing the frequency F of each replica (As shown in Figure 3).

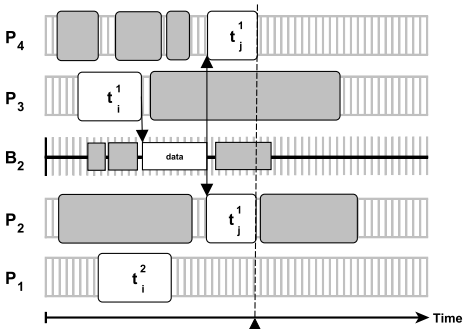


Fig. 3. Changing the frequency of t^1_j

3.2 Passive Redundancy with Variable Data Fragmentation

In order to use efficiently the bus redundancy of the architecture, we propose to use a mechanism of communication, based on variable data fragmentation. Variable data fragmentation allows the fast recovering from buses errors, and it may also reduce the error detection latency. (the time it takes to detect the error). The communication of each data dependency $t_i \rightarrow t_j$ is fragmented into *N+1* fragments $data = data_1 \bullet \cdots \bullet data_{N+1}$, sent by t_i to t_j via *N+1* distinct buses (see Figure 4); The associative operation ($\bullet$) is used to concatenate two data packets. As our approach uses variable data fragmentation, the size of each fragmented data depends on *GSFR* and the bus failure rates λ_B.

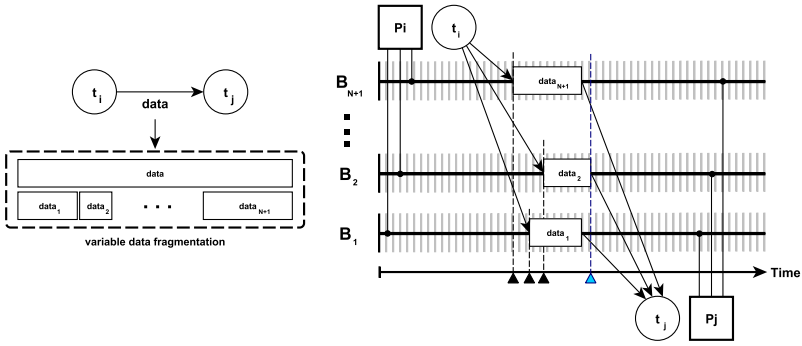


Fig. 4. Variable data fragmentation

GSFR is the failure rate per time unit of the obtained multiprocessor schedule. Using the GSFR is very satisfactory in the area of periodically executed schedules. In such cases, applying brutally the exponential reliability model yields very low reliabilities due to very long execution times (the same remark applies also to very long schedules). Hence, one has to compute beforehand the desired reliability of a single iteration from the global reliability of the system during its full mission; but this computation depends on the total duration of the mission and on the duration of one single iteration.

Our fault tolerance heuristic is GSFR-based to control precisely the scheduling of each fragmented data from the beginning to the end of the schedule. In [16], The GSFR of scheduling an operation t_i, noted $\Lambda(S_n)$, by the following equation:

$$\Lambda(S_n) = \frac{-\log(\prod_i e^{-\lambda_k exe(t_i,P_j)+\sum_k \sum_j \lambda_c exe(dpd_j^k,b_c)})}{\sum_i^j exe(t_i,p_j) + \sum_k^m exe(dpd_k,b_m)} \quad (5)$$

Variable data fragmentation operates in three phases :

1. First, in order to tolerate at most N communication bus errors, each data dependency is fragmented into *N+1* fragments of equal size. The initial size of each fragment is calculated by:

$$Size(data_i) = \frac{Size(data)}{N+1} \tag{6}$$

The main problem with the equal size data fragmentation comes from the difference between ending time of different fragments (Figure 5(a)) because the destination operation must wait to getting all the fragments of the data dependency to start execution.

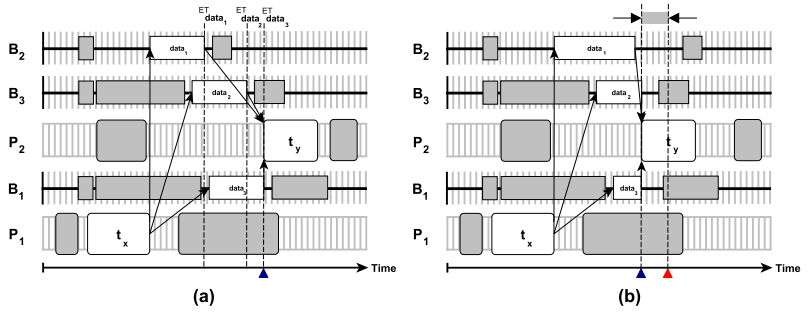


Fig. 5. Ending time : (a) ET in equal size data fragmentation, (b) Minimize difference between ending time

2. Second, the goal of passing from equal size data fragmentation to variable data fragmentation (Figure 5(a)) is to minimize the difference between ending time ET of different fragments (Figure 5(b)).

$$\begin{array}{l} ET_{data_1} \leq ET_{data_2} \leq \cdots \leq ET_{data_{N+1}} \\ Minimize(ET_{data_{i+1}} - ET_{data_i})_{i \in \{1,\cdots,N+1\}} \end{array} \tag{7}$$

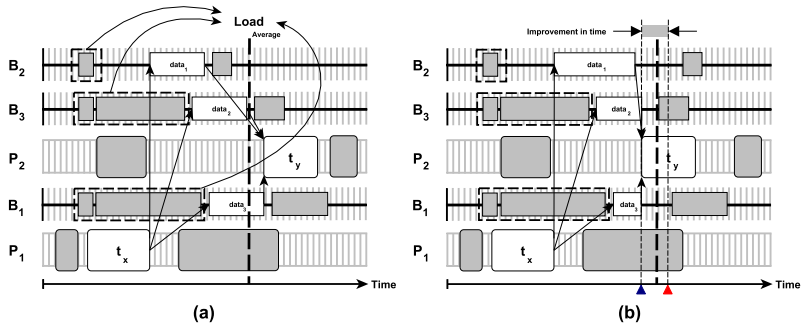


Fig. 6. (a) the Average Load $Load_{Average}$, (b) and the improvement in time of the scheduling

With variable data fragmentation based on minimizing the difference between ending time, another problem can occur and grows extremely the execution time. The bus over which accumulates data may also fail, therefore the quantity of data to be retransmitted is more important.

3. Third, the definition of a compromise between the load of each communication bus and the maximum data to be transmitted on this bus, as illustrated in Figure 6(a). Variable data fragmentation must not exceed this value when defining the new fragments size. The improvement in time of the scheduling is shown in Figure 6(b).

The algorithm that enable variable data fragmentation is show in figure 7.

Algorithm VDF
Input: data-dependence $(data = t_i \rightarrow t_j)$, N.
Output: the set of $N+1$ affectation $(data_i(B_x))$.

1. Each data dependency $(data = t_i \rightarrow t_j)$ is fragmented into $N+1$ fragments of equal size:

$$Size(data_1) = \cdots Size(data_{N+1}) = \frac{Size(data)}{N+1}$$

2. Compute the loading sill of buses.

$$Load_{Average} = \frac{\sum \lambda_{B_i} * Load(B_i)}{N+1}$$

3. Schedule the $N+1$ fragments of data-dependence on $N+1$ bus.
4. Order the data fragments according to their ending Time.

$$ET_1 \leq ET_2 \leq \cdots \leq ET_{N+1}$$

5. Compute the sum of the shift of Ending Time.

$$Sum^{new}_{shift-time} := 0; Sum_{shift-time} = \sum ET_{i+1} - ET_i;$$

6. **While** $(Sum^{new}_{shift-time} \leq Sum_{shift-time})$ **do**
   (a) $Sum_{shift-time} := Sum^{new}_{shift-time}$.
   (b) Fragment the data Fragment with the last end time on tow fragments $(data(ET_{N+1}) = data_A \bullet data_B)$, respecting the following three conditions:
      i. $Size(data_A) \geq Siz_{min}(data_{ET_1})$
      ii. $Siz(data_{ET_1} + Size(data_B) \leq Load_{Average}$
      iii. $ET_1 + Size(data_B)B_{data_1} \leq ET_{N+1}$
   (c) Order the data fragments according to their new ending time ET_i.
   (d) Compute the new value of $Sum^{new}_{shift-time}$

$$Sum^{new}_{shift-time} = \sum ET_{i+1} - ET_i;$$

   End While.

End

Fig. 7. VDF : The variable data fragmentation algorithm

3.3 Scheduling Algorithm

The principles of our approach are implemented by a scheduling algorithm, called Energy Fault Tolerant Heuristic *(EFTH-VDF)*. It is a greedy list scheduling

heuristic, which schedules one operation at each step (n). It generates a distributed static schedule of a given algorithm Alg onto a given architecture Arc, which minimizes the system's run-time, and tolerates upto L processors and N buses faults, with respect to the real-time and the distribution constraints. At each step of the greedy list scheduling heuristic, the pressure schedule function (noted by $\sigma(n)(t_i, P_j)$) is used as a cost function to select the best operation to be scheduled.

$$\sigma^{(n)}(t_i, P_j) = S^{(n)}_{t_i,P_j} + \overline{S}^{(n)}_{t_i} - R^{(n-1)} \tag{8}$$

The *EFTH-VDF* algorithm (show in figure 8) is divided into seven steps.

Algorithm EFTH-VDF
Input: ALG, ARC, N;
Output: a reliable fault-tolerant schedule;

Initialize the lists of candidate and scheduled operations:
n := 0;
$T^{(0)}_{cand} := \{t \in T \mid pred(t) = \emptyset\}$;
$T^{(0)}_{sched} := \emptyset$;

While $(T^{(n)}_{cand} \neq \emptyset)$ **do**

1. For each candidate operation t_{cand}, compute $\sigma^{(n)}$ and GSFR on each processor P_k.
2. For each candidate operation t_{cand}, select the best processor $p^{t_{cand}}_{best}$ which minimizes $\sigma^{(n)}$ and GSFR.
3. Select the most urgent candidate operation t_{urgent} between all t^i_{cand} of $T^{(n)}_{cand}$.
4. For each data dependencies whose t_{urgent} is the producer operation: Fragment the data communication on N fragments using the variable data fragmentation algorithm;
5. Schedule t_{urgent} and its fragmented data;
6. Update the lists of candidate and scheduled operations:
   $T^{(n)}_{sched} := T^{(n-1)}_{sched} \cup \{t_{urgent}\}$;
   $T^{(n+1)}_{cand} := T^{(n)}_{cand} - \{t_{urgent}\} \cup \{t' \in succ(t_{urgent}) \mid pred(t') \subseteq T^{(n)}_{sched}\}$;
7. n := n + 1;

End while
End

Fig. 8. The EFTH-VDF algorithm

4 Simulations, Results and Discussion

We have applied the *EFTH-VDF* heuristic to an example of an algorithm graph and an architecture graph composed of four processors and four buses. The algorithm graph is show in Figure 9. The failure rates of the processors are respectively 10^{-5}, 10^{-5}, 10^{-6} and 10^{-6}, and the failure rate of the Buses SAM_{MP1}, SAM_{MP2}, SAM_{MP3} and SAM_{MP4} are respectively 10^{-6}, 10^{-6}, 10^{-5} and 10^{-4}.

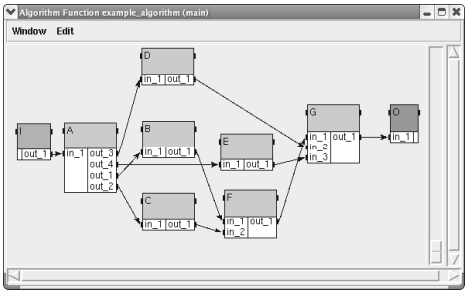


Fig. 9. Algorithm graph

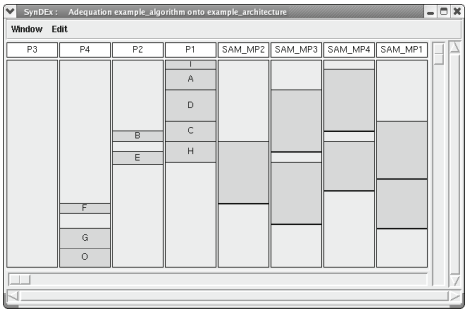


Fig. 10. Schedule generated by SynDEx

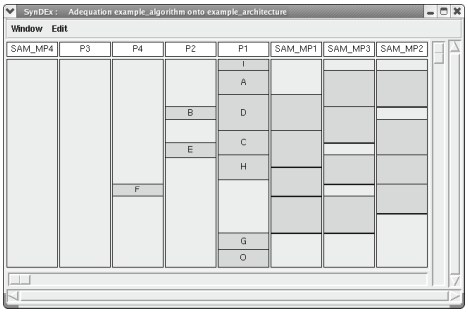


Fig. 11. $EFTH - VDF$ without changing frequencies

Figure 10 shows the non-fault-tolerant schedule produced for our example with a basic scheduling heuristic. (for instance the one of SynDEx). SynDEx [17] is a tool for optimizing the implementation of real-time embedded applications on multi-component architecture.

Figure 11 shows the fault-tolerant schedule produced for our example with a *EFTH-VDF* scheduling heuristic without changing frequencies. The schedule length generated by this heuristic is 21.6. The GSFR of the non-reliable schedule is equal to 0.0000287. The energy E is equal to 36.7.

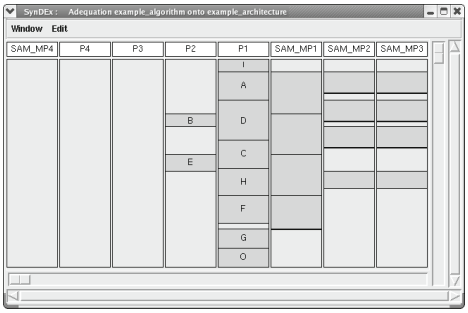


Fig. 12. A schedule generated by $EFTH - VDF$

Figure 12 shows the fault-tolerant schedule produced for our example with a *EFTH-VDF* scheduling heuristic. The schedule length generated by this heuristic is 27.3. The GSFR of the non-reliable schedule is equal to 0.0000276. The energy E is equal to 23.21.

5 Conclusion

We have proposed in this paper a solution to tolerate both processors and communication media faults in distributed heterogeneous architectures with multiple-bus topology. The proposed solution, based on active redundancy, is a list scheduling heuristic called *EFTH-VDF*. It generates automatically distributed static schedule of a given algorithm onto a given architecture, which minimizes the system's run-time, and tolerates upto L processors and N buses faults, with respect to real-time and distribution constraints. The scheduling strategy based on variable frequency and variable data fragmentation minimizes energy consumption and take communication failures into account.

References

1. Jalote, P.: Fault-Tolerance in Distributed Systems. Prentice Hall, Englewood Cliffs (1994)
2. Kopetz, H.: Real-time systems: design principles for distributed embedded applications. Springer Science & Business Media (2011)
3. Grünsteidl, G., Kantz, H., Kopetz, H.: Communication reliability in distributed real-time systems. In: Distributed Computer Control Systems 1991: Towards Distributed Real-Time Systems with Predictable Timing Properties, p. 123 (2014)
4. Jun, Z., Sha, E.H., Zhuge, Q., Yi, J., Wu, K.: Efficient fault-tolerant scheduling on multiprocessor systems via replication and deallocation. International Journal of Embedded Systems 6(2), 216–224 (2014)
5. Kandasamy, N., Hayes, J.P., Murray, B.T.: Dependable communication synthesis for distributed embedded systems. Reliability Engineering & System Safety 89(1), 81–92 (2005)

6. Dulman, S., Nieberg, T., Wu, J., Havinga, P.: Trade-off between traffic overhead and reliability in multipath routing for wireless sensor networks. In: Wireless Communications and Networking Conference (2003)
7. Dima, C., Girault, A., Lavarenne, C., Sorel, Y.: Off-line real-time fault-tolerant scheduling. In: 9th Euromicro Workshop on Parallel and Distributed Processing, pp. 410–417 (2001)
8. Pinello, C., Vincentelli, L.C., Fault-tolerant, A.S.: deployment of embedded software for cost-sensitive real-time feedback-control applications design. In: Automation and Test in Europe, DATE 2004. IEEE (2004)
9. Song, H., Wu, H.: The applied research of support vector machine in bus fault identification. In: 2010 Sixth International Conference on Natural Computation (ICNC), vol. 3, pp. 1326–1329. IEEE (2010)
10. Izosimov, V., Pop, P., Eles, P., Peng, Z.: Scheduling and optimization of fault-tolerant embedded systems with transparency/performance trade-offs. ACM Transactions on Embedded Computing Systems (TECS) 11(3), 61 (2012)
11. Krishna, C.: Fault-tolerant scheduling in homogeneous real-time systems. ACM Computing Surveys (CSUR) 46(4), 48 (2014)
12. Huang, J., Buckl, C., Raabe, A., Knoll, A.: Energy-aware task allocation for network-on-chip based heterogeneous multiprocessor systems. In: 2011 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), pp. 447–454. IEEE (2011)
13. Agrawal, P., Rao, S.: Energy-aware scheduling of distributed systems using cellular automata. In: 2012 IEEE International Systems Conference (SysCon), pp. 1–6. IEEE (2012)
14. Agrawal, P., Rao, S.: Energy-aware scheduling of distributed systems. IEEE (2014)
15. Zhu, D., Melhem, R., Mosse, D., Elnozahy, E.: Analysis of an energy efficient optimistic tmr scheme. In: Proceedings of the Tenth International Conference on Parallel and Distributed Systems, ICPADS 2004, pp. 559–568. IEEE (2004)
16. Girault, A., Kalla, H.: A novel bicriteria scheduling heuristics providing a guaranteed global system failure rate. IEEE Transactions on Dependable and Secure Computing 6(4), 241–254 (2009)
17. Forget, J., Gensoul, C., Guesdon, M., Lavarenne, C., Macabiau, C., Sorel, Y., Stentzel, C.: Syndex v7 user manual (2013)

Genetic Centralized Dynamic Clustering in Wireless Sensor Networks

Mekkaoui Kheireddine[1(✉)], Rahmoun Abdellatif[2], and Gianluigi Ferrari[3]

[1] GeCoDe Laoratory, University of Dr Tahar Moulay, Saida, Algeria
[2] EEDIS Laboratoty, University of Djilllai Liabes, SBA, Algeria
[3] WASNLab Laboratory, University of Parma, Parma, Italy
mekdar@hotmail.com

Abstract. In order to overcome the energy loss involved by communications in wireless sensor networks (WSN), the use of clustering has proven to be effective. In this paper, we proposed a dynamic centralized genetic algorithm (GA)-based clustering approach to optimize the clustering configuration (cluster heads and cluster members) to limit node energy consumption. The obtained simulation results show that the proposed technique overcomes the LEACH clustering algorithm.

1 Introduction

Wireless sensor networks (WSNs) are used in many domains, such as military surveillance, disaster management, forest fire detection, seismic detection, habitat monitoring, biomedical health monitoring, inventory tracking, animal tracking, hazardous environment sensing and smart spaces, general engineering, commercial applications, home applications, underwater applications, etc [1]. Indeed, according to [2], WSNs are considered to be one of the new technologies that will change our life, they are listed, also in [3], as one of the key technologies of the internet of things.

The sensor nodes (or motes) are physical entities characterized by: (i) a battery with a limited energy; (ii) a processor with limited processing capabilities; (iii) and a transceiver [4]. The nodes can be deployed in monitoring areas in order to gather multiple types of information (e.g., humidity, light, temperature, wind,...) and then transmit the gathered information to the gateway sensor node (Access Point or Sink), possibly using multi-hop routing strategy [5]. In turn, the sink transmits the collected information to the end users.

Since it might often be difficult to replace exhausted batteries (e.g., WSNs may be deployed in inaccessible areas) [6], extending the lifetime of the WSN is crucial. In the literature, many papers show that the source of highest energy consumption in the sensor node is the transceiver [7], making strategies which minimize the use of the transceiver very attractive. Several techniques can be used to save energy, among which clustering consists in grouping sensors in several clusters, so that each cluster has a single cluster-head and several cluster-members. In each cluster, the cluster-members gather information on the sensed area and send it to the cluster-head. In turns, the cluster-head processes the data

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 503–511, 2015.
DOI: 10.1007/978-3-319-19578-0_41

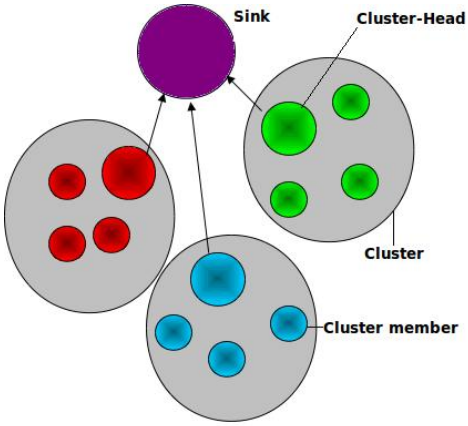


Fig. 1. Clustered WSN

received from its members and sends it to the sink. A graphical representation of a clustered WSN is shown in Fig. 1.

In a clustered WSN, data collected by the sensors is communicated to its cluster-head, for data processing and redundancy elimination. Therefore, sensors communicate data over short distances in each cluster (to cluster-heads), so that the energy spent in communication will be lower than that spent with sensors communicating directly to the sink [8].

Clustering can be static or dynamic. In a *static* scenario, the cluster-heads are fixed and tend to exhaust their energies rapidly, making this clustering unsuitable for WSNs [9]. In fact, the network becomes nonfunctional in the absence of cluster-heads. In the presence of *dynamic* clustering, the clusters change over the time, equalizing the energy consumption across all nodes and, thus, extending the network lifetime.

In this paper, our goal is to maximize network lifetime (defined as the time interval from the nodes' deployment to the instant at which a given percentage of deployed nodes die [6]) by minimizing the average energy consumption of all nodes. In order to do this, all nodes can be promoted to the role of cluster-heads. In order to reach this goal and guarantee full coverage (i.e., the clusters are spread over the entire network), we rely of on the use of a genetic algorithm (GA), which determines, in each cycle, wether or not a node can be chosen to play the role of a cluster-head.

2 Related Work

The idea of using clustering has been adopted by many authors. The linked clustering algorithm (LCA) was one of the first approaches [10]. In the LCA

algorithm, each node has a unique ID. In this algorithm a node play the role of cluster-head if its ID is the highest one in its neighboring.

LEACH is the most popular clustering algorithm for WSNs [9]. LEACH allows a fixed percentage of nodes to become cluster-heads (namely, 5% of the nodes) and leads to the creation of clusters in a distributed way, with the nodes taking autonomous decisions. Each node decides to become a cluster-head with probability p. A node which does not become cluster-heads determines its cluster by choosing the nearest cluster-head. On average, LEACH provides low energy consumption and a uniform energy distribution among the nodes. However LEACH has also some drawbacks. Because of the probabilistic selection of the cluster-heads, a node with a very low energy can be selected as cluster-head. Moreover, since the selection of cluster-heads is probabilistic, the chosen cluster-heads may be placed in the same area, so that a good coverage can not be guaranteed: in fact, some nodes will be disconnected from the network (i.e., they will not attach to any cluster-head). Moreover, the use of a fixed percentage of cluster-heads may lead (network-wide) to higher energy consumption, as the number of cluster-heads depend on several factors, such as node spatial density [11].

EEHC is a randomized and distributed clustering algorithm, whose goal is to maximize the network lifetime [12]. This algorithm is executed in two levels. In the first level, denoted as "initial," volunteer nodes, which do not belong to any cluster, may decide to be cluster-heads with probability p and they announce their decisions to their neighbors. The nodes that do not receive an announcement, within a specified time interval t, become forced cluster-heads. In the second level, denoted as "extended," the clustering algorithm is recursively repeated to form hierarchical clustering, where new cluster-heads are selected from the already formed cluster-heads, until a final base station is reached.

In [13], the authors consider a GA and adapt, on the basis of software services, its parameters to determine the energy consumption and, therefore, extend the network lifetime. In [6], the authors proposed a GA-inspired routing protocol (GROUP): in particular, they use GA and simulated annealing (instead of the greedy chain) to select routing paths efficiently.

3 System Model

The conditions and assumptions behind the considered network model are compliant with those considered in [9] for LEACH. More precisely, they can be summarized as follows.

1. The base station is fixed, is not energy-constrained, and has a high computing capacity.
2. All the nodes deployed are energy/power-constrained and homogeneous.
3. The data processing power is very low with respect to the power required to transmit and receive data.

The nodes' radio communication specifications are set as in [9,4,6]. In particular, we assume that the radio module dissipates: $E_{elec} = 50$ nj/bit in transmission/receiver circuitry; and $\epsilon_{amp} = 100$ pJ/bit/m^2 in the transmitter amplifier.

Considering free space communications, in order to transmit a k-bit message over a distance d (dimension: [m]) a node consumes the following amount of energy:

$$E_{Tx}(k,d) = E_{elec} \times k + \epsilon_{amp} \times k \times d^2. \quad (1)$$

When receving a k-bit message a node consumes the following amount of energy:

$$E_{Rx}(k) = E_{elec} \times k. \quad (2)$$

4 The Proposed Approach

4.1 The Problem

In a clustered WSN, if a few cluster-heads are used, then most of the nodes are likely to have a long transmit radio range to send the collected data to their cluster-heads and this tends to quickly deplete their batteries' energies. If a large number of cluster-heads is used, this leads mostly to a one-hop network (most nodes are cluster-heads and must reach the base station in one hop): this consumes also quickly the battery energy [4,8].

The best clustering strategy consists in optimizing (i) the number of cluster-heads and (ii) their positions. In particular, a node can be promoted to cluster-head according to several parameters: its residual energy, its distance to the sink, and the sum of the distances to its cluster-members. This suggests the use of GAs to find the optimal combination of these parameters.

4.2 The Proposed Algorithm

In this paper, we consider *dynamic* clustering, i.e., re-clustering is considered to avoid early death of cluster-heads. The proposed GA is executed at the sink (i.e., it is *centralized*), due to the needed computing capacity, and the obtained results (in terms of clustering configuration) are communicated to the nodes. At each re-clustering round, each node can then be either a cluster-head or a cluster-member. This centralized approach is expected to overcome the main limitations of LEACH, where the number of cluster-heads is fixed and their spatial distribution is arbitrary, i.e,, there is no coordination [9].

In order to use a GA, a WSN needs to be "codified." In particular, we use a binary representation, in which each node is represented either by 0 (if it is a cluster-member) or by 1 (if it is a cluster-head). Each codified network is called a "chromosome." A set of chromosomes is called a "generation."

The used GA is based on exploration and exploitation of the entire research space using an evolutionary strategy, it helps us to find an optimal combination of cluster-heads, cluster-members and their distributions in the monitoring area, among many combinations existing in the research space, making the energy consumption and the network coverage, optimal. Each potential solution is characterized by a value called fitness, which determines the optimality of

solutions. In correspondence to a generation, the GA keeps the best chromosomes and drops others according to their fitness function. Each chromosome, in fact, represents a potential solution. The GA then applies the following genetic operators to generate new offsprings [14].

- *Selection.* The selection process is used to choose the best chromosomes from a generation. In our simulation, the roulette wheel algorithm is used to perform the selection.
- *Crossover.* To apply crossover, we choose arbitrary two chromosomes from a generation, we choose, also, two random positions in the chosen chromosomes and we used the two point crossover, to generate two new offsprings, that will belong to the next generation.
- *Mutation.* The mutation is used to avoid the super chromosome problem. It means if one chromosome is selected many times in the same generation, the crossover will not produce new chromosomes, since the parents are the same chromosome. Hence, the mutation is used to change, in each chromosome, an arbitrary bit. Several tentatives have been performed so to come up with the best-run GA parameters in terms of runtime and convergence. The best crossover and mutation probabilities are 0.75 and 0.2, respectively.

As mentioned above, each chromosome is then evaluated with a fitness function which attributes a higher chance to the best solutions to survive. The fitness of a candidate chromosome can be expressed as follows:

$$\text{Fitness} = f(NNN, NCH, DNCH, RECH) \tag{3}$$

where:

- NNN is the number of networked nodes;
- NCH is the number of cluster-heads;
- $DNCH$ is the sum of the distances between the cluster-members (CMs) and their cluster-heads (CHs), i.e.,

$$DNCH = \sum_{i \in \{\text{CHs}\}} \sum_{j \in \{\text{CMs}\}} \text{Distance}(\text{CH}_i, \text{CM}_j);$$

- $RECH$ is the sum of residual (cumulative) energy at the cluster-heads (dimension: [mW]), i.e.,

$$RECH = \sum_{i \in \{\text{CHs}\}} \text{Residual energy at the CH}_i\,.$$

In order to optimize the proposed clustering mechanism, we consider the following (heuristic) fitness function:

$$\text{Fitness} = (NNN)^{\alpha_1} + \left(\frac{NNN}{NCH}\right)^{\alpha_2} + (10^3 \cdot RECH)^{\alpha_3} + DNCH^{\alpha_4}$$

where: the exponential parameters $\{\alpha_i\}_{i=1}^4$ need to be properly optimized; the fraction NNN/NCH represents the average cluster dimension; the multiplicative term 10^3 used for $RECH$ properly weighs the energy dimension. By trial and error, the best fitness function (i.e., the best configuration of the exponents $\{\alpha_i\}_{i=1}^4$) turns out to be

$$\text{Fitness} = (NNN)^6 + \left(\frac{NNN}{NCH}\right)^5 + (10^3 \cdot RECH)^2 + DNCH.$$

5 Performance Analysis

In order to validate the proposed clustering approach, we carry out a simulation-based performance analysis, considering different scenarios, by varying the node spatial density, the number of nodes, and the sink position. In each scenario, a given number of sensors is randomly deployed in a square monitored area, with side length 800×600. The sink, placed within the region, runs the GCDC algorithm (*for Genetic Centralized Dynamic Clustering*) and informs the sensors of the decided clustered configuration. After receiving the decision, each node knows if it is a CH or a cluster member. The GCDC algorithm is periodically run by the sink in order to avoid that a node death compromises network connectivity. We assume that all nodes have batteries with initial energy equal to 0.25 J. The dimension k of the messages to be transmitted is set to 100 bits. We assume that random "events" (e.g., acoustic signal detection, motion detection, etc.) happen in the monitored area: in particular, each random event is detected by its nearest neighbor, which needs to report this observation to the sink. In all considered scenarios, the performance of the GDC algorithm is compared with that of LEACH. Two values of the initial number of nodes in the WSN are considered: 100 (low node spatial density) and 1000 (high node spatial density).

In Figure 2, the residual network energy is shown as a function of the simulation time (expressed in event number), considering two values for the initial number of nodes: (a) 100 and (b) 1000. It can be observed that GCDC algorithm allows to save more energy than LEACH. The energy saving is not relevant at the beginning, whereas it becomes more significant as the time passes by. This is due to the fact that the GCDC algorithm updates the network clustered topology very efficiently. This behaviour is more pronounced in the scenario with 100 nodes (low node spatial density).

In Figure 3, we investigate the network connectivity evolution, considering (a) NNN (i.e., the network coverage) and (b) the number of dead nodes, as a function of the simulation time (in terms of event number). In both cases, the initial number of nodes in the WSN is set to 100 (low node spatial density). From the results in Figure 3 (a), it can be observed that the number of nodes connected (i.e., becoming cluster members or heads) by the GCDC algorithm is larger than that guaranteed by LEACH. This is more evident at the beginning of the simulation, when all nodes (having full battery energies) could be

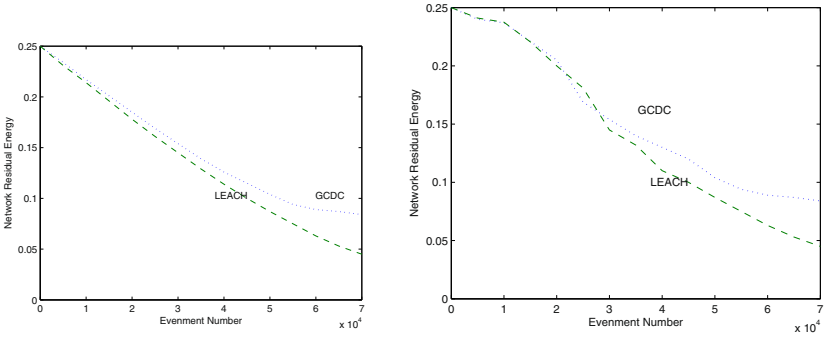


Fig. 2. Network residual energy as a function of the simulation time (in terms of event number). The initial number of nodes in the network is set to: (a) 100 or (b) 1000.

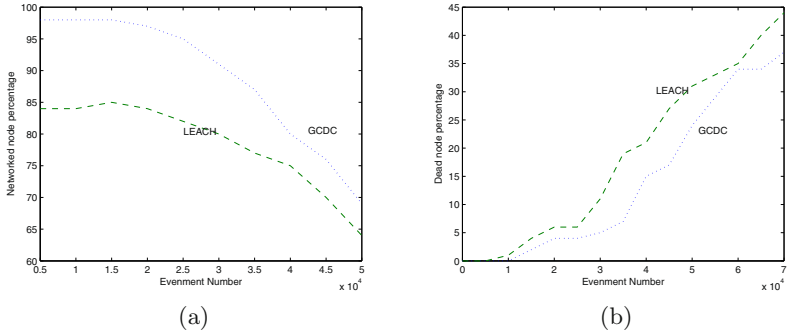
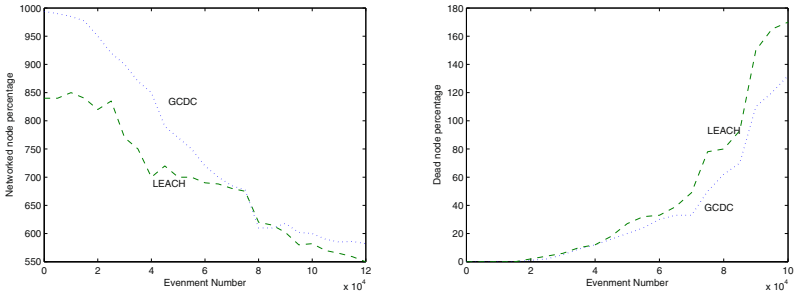


Fig. 3. Network connectivity evolution, in terms of (a) NNN and (b) number of dead nodes, as a function of the simulation time (in terms of event number). The initial number of nodes in the WSN is 100.

connected, whereas the improvement brought by GCDC reduces for advancing simulation time, as a larger and larger number of nodes die. The performance difference is due to the fact that LEACH *a priori* sets the number of CHs to 5% of the total number of nodes without identifying their positions: this likely leads to overlapped clusters (i.e., two CHs may be close to each other), leaving other nodes (without a sufficiently close CH) disconnected. The GCDC algorithm does not determine a priori the number of CHs but, rather, the GA determines the optimized number of CHs, along with their positions, to cover the entired monitored area efficiently. In Figure 3 (b), the number of dead nodes (after energy depletion) is shown: as expected from Figure 3 (a), the death rate with GCDC is lower than that with LEACH, owing to the clustering procedure which takes into account the nodes' residual energies.

In Figure 4, the network connectivity evolution, considering (a) NNN (i.e., the network coverage) and (b) the number of dead nodes as functions of the



(a) (b)

Fig. 4. Network connectivity evolution, in terms of (a) NNN and (b) number of dead nodes, as a function of the simulation time (in terms of event number). The initial number of nodes in the WSN is 1000.

simulation time (in terms of event number), is investigated in a scenario with 1000 initial nodes (high node spatial density). By comparing the results in Figure 4 (a) with those in Figure 3 (a), it can be concluded that the performance improvement, in terms of NNN, brought by GCDC is more pronounced in dense network. In particular, since, according to the results in Figure 4 (b), the death rates of GCDC and LEACH are approximately the same, it means that the GCDC is very efficient in reclustering the topology in order to guarantee a high level of connectivity to the surviving nodes.

6 Conclusion

In this paper, we have presented a novel clustering algorithm, denoted as GCDC, which uses a GA to optimize the number and the corresonding locations of CHs. The performance of our algorithm has been compared with that of LEACH. The obtained results show that the proposed clustering algorithm reduces the (network-wide) energy depletion rate and guarantees a better network coverage. An interesting research direction consists in applying the proposed GA-based clustering algorithm to duty-cycled WSNs.

References

1. Huafeng, W., et al.: An acoa-afsa fusion routing algorithm for underwater wireless sensor network. International Journal of Distributed Sensor Networks, 4110–4118 (2012)
2. Ilyas, M., Mahgoub, I.: Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems. CRC Press LCC (2012)
3. Jianbin, X., Ting, Z., Yan, Y., Wenhua, W., Songbai, L.: Cooperation-based ant-colony algorithm in wsn. Journal of Networks 8(4) (2013)

4. Mekkaoui, K., Rahmoun, A.: Short-hops vs. long-hops - energy efficiency analysis in wireless sensor networks. In: CIIA 2011:Proceedings of the Third International Conference on Computer Science and its Applications (CIIA11), University of Saida, Algeria, pp. 13–15 (2011)
5. Akyildiz, I.F., Vuran, M.C.: Wireless Sensor Networks, Jon S.Wilson
6. Chakraborty, A., Mitra, S.K., Naskar, M.K.: A genetic algorithm inspired routing protocol for wireless sensor networks. International Journal of Computational Intelligence Theory and Practice 6(1) (2011)
7. Odey, A.J., Li, D.: Low power transceiver design parameters for wireless sensor networks. Wireless Sensor Network 4(10), 243–249 (2012)
8. Abbasi, A.A., Younis, M.: A survey on clustering algorithms for wireless sensor networks. Computer Communications 30(14), 2826–2841 (2007)
9. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, HICSS 2000, p. 8020. IEEE Computer Society, Washington, DC (2000)
10. Dechene, D.J., Jardali, A.E., Luccini, M., Sauer, A.: Wireless sensor networks a survey of clustering algorithms for wireless sensor networks, Department of Electrical and Computer Engineering, The University of Western Ontario, Canada, Tech. Rep (2006)
11. Jin, S., Zhou, M., Wu, A.S.: Sensor network optimization using agenetic algorithm. In: Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics, Orlando, FL, pp. 109–116 (2003)
12. Bandyopadhyay, S., Coyle, E.J.: An energy efficient hierarchical clustering algorithm for wireless sensor networks. In: Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, INFOCOM 2003, pp. 1713–1723 (2003)
13. Norouzi, A., Babamir, F.S., Zaim, A.H.: 'A New Clustering Protocol for Wireless Sensor Networks Using Genetic Algorithm Approach. Wireless Sensor Network 3(11), 362–370 (2011)
14. Holland, J.H.: Genetic algorithms. Scientific American 267(1), 66–72 (1992)

Security and Network Technologies: Potpourri

Region-Edge Cooperation for Image Segmentation Using Game Theory

Omar Boudraa(✉) and Karima Benatchba

Doctoral School (STIC), High School of Computer Sciences (ESI),
Oued Smar, Algiers, Algeria
{o_boudraa,k_benatchba}@esi.dz

Abstract. Image segmentation is a central problem in image analysis. It consists of extracting objects from an image and separating between the background and the regions of interest. In the literature, there are mainly two dual approaches, namely the region-based segmentation and the edge-based segmentation. In this article, we propose to take advantage of Game theory in image segmentation by results fusion. Thus, the presented game is cooperative in a way that both players represented by the two segmentation modules (region-based and edge-based) try coalitionary to enhance the value of a common characteristic function. This is a variant of the parallel decision-making procedure based on Game theory proposed by Chakraborty and Duncan [1]. The involved pixels are those generated from the cooperation by results fusion between the edge detector (Active contour) and the region detector (Region growing) posing a decision-making problem. Adding or removing a pixel (to/from) the region of interest depends strongly on the value of the characteristic function. Then, and to study the effectiveness and noise robustness of our approach we proposed to generalize our experimentations, by applying this technique on a variety of images of different types taken mainly from two known test databases.

Keywords: Region-based segmentation · Edge-based segmentation · Region-edge cooperation · Game theory · Nash equilibrium

1 Introduction

Image segmentation plays a key role in image analysis. In addition, it determines the quality of characteristics measures calculated later in image understanding process. However, there are mainly two dual approaches of segmentation. The edge-based segmentation approach that locates the boundaries of objects; and the region-based segmentation approach which partitions the image into a set of regions. Each region defines one or more connected objects.

In order to improve the results of each approach by trying to combine their own advantages, researchers have created what is called cooperative segmentation [2].

Game theory is a strong tool for analyzing situations, modeling and determining the best strategy(ies), often used in Economics and in a variety of domains. This theory proves interesting in this case given the principle of duality region-edge and the problem of antagonism between the two image segmentation approaches.

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 515–526, 2015.
DOI: 10.1007/978-3-319-19578-0_42

In our contribution, we propose to take advantage of Game theory in image segmentation by results fusion. It is to treat both types of segmentation, in a coalitionary way as two players exchanging information in "*Game Theory Integrator*" module to simultaneously improve their individual results.

This article consists of three sections. The first section presents general information on individual and cooperative techniques of image segmentation, its different forms and a bibliographical study on the integration of Game theory in image segmentation and its contribution. The second section details our contribution. While, the last section is devoted to experimentations, results and evaluation of the performance of our approach and its robustness to noise.

2 Around Image Segmentation and Game Theory

2.1 Image Segmentation

Segmentation is the partition of an image into a set of distinct regions (which do not overlap) and whose union is the whole image [3].

2.2 Image Segmentation Approaches

Image segmentation methods can be divided essentially into two categories which are based on two properties between neighboring pixels: *discontinuity* and *similarity*. The *discontinuity* is used by edge-based segmentation approach (boundary), while the *similarity* of pixels is used by region-based segmentation approach.

Edge Approach

The edge approach tries to identify changes between regions. In general, an edge element is a point of the image belonging to the boundary between two or more objects having different grayscale levels.

Derivate methods

The derivate methods are most used to detect the pixels intensity transitions [4]. Overall, they can be classified into two big categories: *Gradient* approach that uses the first derivative and *Laplacian* approach that uses the second derivative.

Deformable models

Segmentation algorithms based on deformable models have the advantage, compared to derivate methods that provide closed edges or surfaces [5]. These methods include: *Active contours* and *Level sets.*

Limitations of the edge-based segmentation

Edge-based segmentation has some limitations and drawbacks such as the difficulty of identification and classification of parasite edges. In addition, the detected edges are not always closed. Nevertheless, the major weakness is that the edge-based segmentation does not give comprehensive information on the content of the image [6].

Region Approach

This approach consists in dividing the image into distinct regions [7]. In contrast to the edge approach, there methods are interested in the region content. The most common techniques for region-based segmentation are shown in the following.

Region growing

Region growing technique is based primarily on the notion of seed. A seed is one pixel or set of pixels (region). From it, regions are constructed by aggregation of adjacent and homogeneous pixels (grayscale, color similarity...etc.) [7]. The Region growing process stops when all pixels have been processed (assigned to a region).

This technique is simple and quick to perform. In addition, it allows the object segmentation in complex topology [8]. Whereas, the choice of initial germs and homogeneity criterion is critical.

Region Splitting

Region splitting technique involves image partitioning into homogeneous regions according to a given criterion. Its principle is to consider the image as the initial region, which then is divided into regions. The splitting process is repeated for each new region until homogeneous regions [9]. Its drawback is the *over-segmentation*.

Region Merging

Region merging technique is a bottom-up method. Initially, each image pixel is considered as an elementary region. The method tents gradually to merge the related regions that satisfy a given predicate *P* [9].The process is repeated until the satisfaction of a stopping criterion (usually the visiting of the entire image) [10]. However, this method can introduce the *sub-segmentation* effect.

Region Splitting and Merging

It is a hybrid method, in which, a splitting step is performed first. Its result is injected to the second process (merging similar regions) that corrects the possible effect of over-segmentation introduced by the splitting process.

Limitation of region-based segmentation

Region-based segmentation has some disadvantages that we present below:

- The obtained regions do not always correspond to the objects in the image.
- The limits of the obtained regions are generally imprecise.
- The difficulty of identifying criteria for pixels aggregation or regions division.

Cooperative Approach

As we have seen previously, the region and the edge approaches have both advantages and disadvantages. Researchers have tried to take benefits from the strengths of both approaches and duality concepts between them and gave rise to what is called the cooperative segmentation. It combines the advantages of both solutions: *precision* and *speed* of edge-based segmentation, *boundary closures* and *density* of the extracted information of region-based segmentation [2].

Depending how to cooperate the both processes, the researchers proposed three different approaches: *Sequential cooperation*, *fusion results cooperation* and *mutual cooperation*.

Sequential cooperation
The general principle of the sequential cooperation is one of the individual techniques is executed first. Its result is then exploited by the second technique [11].

Fusion results cooperation
In fusion results cooperation, region-based and edge-based segmentation are executed in parallel and independently. Cooperation takes place at their respective results.

Mutual cooperation
In mutual cooperation approach, different segmentation techniques are executed in parallel while mutually exchanging information.

2.3 About Game Theory

Game theory is a formalism that aims to study the *planned*, *real* or *posteriori* justified behavior of agents deal with situations of *antagonism* (opposition), and seek to highlight *optimal* strategies [12]. It is based on the concept of game defined by a set of players (considered as rational agents), all the possible strategies for each player, and the gains specification of players for each combination of strategies [13].

Types of Games
The most popular types of games are:

- *Cooperative and non-cooperative games.*
- *Finite and infinite games.*
- *Synchronous and asynchronous games.*
- *Zero-sum games and non-zero-sum games.*
- *Complete information games and perfect information games.*

Nash Equilibrium
In 1950, *John Nash* has defined a stable interaction situation if no player has interest to change its strategy knowing strategies of others. The game becomes stable that no player can only change its strategy without weakening his own position [14].

Theoretically, it is said that a combination of strategies $s*$ is a Nash equilibrium if the following inequality is satisfied for each player i [14].

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*). \forall\, s_i \in S_i \quad (1)$$

More Clearly, if player i anticipates that the other players will choose the strategies associated with the combination of strategies s_{-i}^*, it can only maximize its gain u by choosing the strategy s_i^*.

2.4 Image Segmentation and Game Theory

Works on the matching between Game theory and image segmentation are not numerous. One possible reason is that Game theory is based primarily to satisfy economic needs. Whereas, the first published work is that of A. CHAKRABORTY et al. in 1999 [15, 1]. In this section, we will quote it with other work in this domain.

- Work of (A. CHAKRABORTY et al. in 1999) is an original and outstanding work that is based on a solid mathematical model integrating Game theory in image segmentation by mutual cooperation between the edge detector (Active contour) and the region detector (Markov Random Fields). It represents a reference work.
- (E. Cassel et al. in 2007) proposed a modified and simplified implementation of Chakraborty and Duncan approach [1]. This simplification involves removing the "Prior information about the form to segment" in the equation of the edge detector. The authors in [16] opted for the "Region growing" as region detector and the morphological operation "closure" for the edge detector.
- Even, (K. ROY et al. in 2010) have proposed an approach to iris and pupil segmentation based on Chakraborty and Duncan work [1]. However, this approach is suitable particularly on this special field of application. For this, they integrated pre-treatments and post-treatments phases in their procedure. In this work, the "Region growing" and "Level sets" methods were used. [17]
- The last two works consist of two individual segmentation approaches (edge-based segmentation only). (B. IBRAGIMOV et al. in 2011) proposed a supervised algorithm based on Game theory and dynamic programming for the segmentation of lung fields [18], while (M. KALLEL et al. in 2013) proposed an approach based on Game theory to restore and segment simultaneously noisy images [19].

3 Cooperative Segmentation Approach Using Game Theory

Now, we present our approach. We propose segmentation by results fusion, suggesting a cooperative game where both players, represented by the two segmentation modules, try coalitionary to improve the value of a common characteristic function. This is mainly based on the work done by Chakraborty and Duncan [1]. This choice is based on the fact that their procedure is original, robust and has been proven mathematically. Indeed, Cassel et al [16] and Roy et al [17] works gives us the opportunity to suggest improvements and changes in the cost functions of this procedure.

3.1 Game Formulation

Now, we define our game and detail its constituting elements, its type and nature.

Game Components
Following Chakraborty and Duncan procedure [1], the objective functions are:
For the *region-based segmentation module* (player 1),

$$F^1(p^1\ ,p^2) = min_X\left[\sum_{i,j}[y_{i,j} - x_{i,j}]^2 + \lambda^2\left(\sum_{i,j}(x_{i,j} - x_{i-1,j})^2 + \sum_{i,j}(x_{i,j} - x_{i,j+1})^2\right)\right] + \acute{a}\left[\sum_{(i,j)\in A_p}(x_{i,j} - u)^2 + \sum_{(i,j)\in\overline{A_p}}(x_{i,j} - v)^2\right]. \quad (2)$$

Where:

- $A_{\vec{p}}$ Corresponds to the set of points which lie inside the contour vector $\vec{p}$, while $\overline{A_{\vec{p}}}$ correspond to the points that lie outside it. Thus, $A_{\vec{p}} \cup \overline{A_{\vec{p}}} = \{(i, j)\ ;\ 1 \le i \le M\ , 1 \le j \le N\}$ = Whole image.
- $u_{i,j}$ represents the information concerning the intensities of points inside the contour and $v_{i,j}$ for points outside.

Also, y is the intensity of the original image, x is the segmented image provided by p^1, u and v corresponds to the intensity mean value of the image on the inside (outside respectively) of the contour given by p^2. The first term attempts to minimize the difference between the values of the pixels intensities found in the region and to strengthen continuity. Whereas, the second term is trying to match between the region and the detected contour.

Whereas, the objective function of player 2 (*edge-based segmentation module*) is:

$$F^2(p^1\ ,p^2) = arg\ \max_{\vec{p}}\left[M_{gradient}(I_g, \vec{p}) + \hat{a}M_{region}(I_r, \vec{p})\right] \quad (3)$$

Where $\vec{p}$ denotes the contour parameterization proposed by p^2, I_g is the gradient image, and I_r is the segmented region obtained by p^1. $M_{gradient}$ represents a correspondence measure (matching) between the gradient image I_g and the detected contour. While, M_{region} is a matching measure between the segmented region image I_r and the contour vector $\vec{p}$, β is its weight.

In our approach, we propose a new formula simplifying function F^2 by replacing:

- $M_{gradient}$ and M_{region} by Abdou and Pratt measure [4].
- Contour parameterization $\vec{p}$ by the constituent pixels of the Active contour.
- The gradient image I_g by *Canny* detector.
- The image of the segmented region I_r only by its boundary.

Finally, we proposed to unify the two cost functions above in one function F:

$$F = \frac{F_{i-1}^1(p^1,p^2) - F_i^1(p^1,p^2)}{F_{i-1}^1(p^1,p^2)} + \frac{F_i^2(p^1,p^2) - F_{i-1}^2(p^1,p^2)}{F_{i-1}^2(p^1,p^2)} \quad (4)$$

Adding or removing a pixel (to/from) the region of interest depends strongly on the improvement or deterioration of this function value. F_i^1 and F_i^2 represent the cost functions of the two segmentation modules, the index i determines whether the pixel i is taken into account or not (i-1 for no and i for yes).

Not only it takes into account the two cost functions, it also helps to normalize the rate of improvement or deterioration of each function because the variation of the function F^1 is almost always greater than those of F^2 as both are not commensurable.

Game Type

By inference, the proposed game is a: *Finite*, *Cooperative* and *Non-zero-sum* game.

3.2 Architecture of the Adopted Approach

We can summarize the organization of our system through a series of interactive modules for segmentation of an image. Each one is presented in the following:

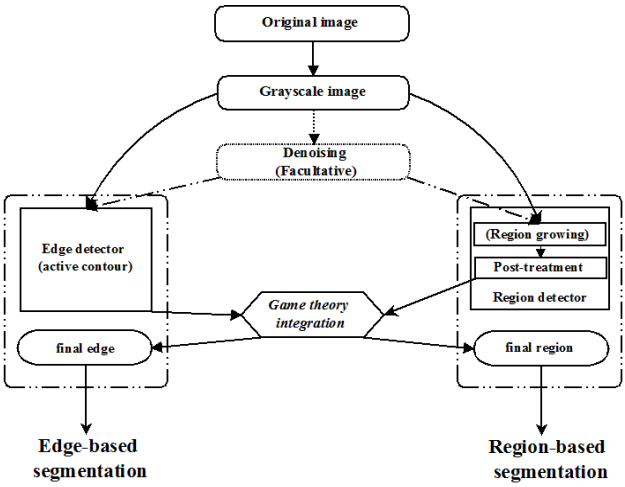


Table 1. Fig. **1.** Overview of the proposed system

Grayscale Conversion
This pretreatment is designed to simplify the image which makes easier the application of our procedures and comparisons; by reducing the amount of information.

Denoising
It consists of an optional treatment, serving to smooth the image (blur effect), reduces noise (unwanted signals) and reduces detail in order to improve the image quality.

Region Detector
As a region detector module, we opted for the Region growing technique. The growing is through the aggregation of candidates' pixels similar to the initial germ of the region, while seeking to minimize the following cost function [16]:

$$E = \sum_{i,j}\left(y_{i,j} - x_{i,j}\right)^2 + \ddot{e}^2 \left(\sum_{i,j}\sum_{i_s,j_s}\left(x_{i,j} - x_{i_s,j_s}\right)^2\right) \quad (5)$$

Where i_s and j_s are the pixel neighborhood of the indices $x_{i,j}$ (in the classification image). The classification image (or segmentation) is initialized by the pixel intensity values in the image processed by the famous Otsu's thresholding technique [20]. That is applied to the grayscale image in order to overcome the problem of non-initialization of neighboring pixels that's not yet been processed.

In our case, and in order to reduce the number of calculations, we proposed at each iteration an estimation of the formula (5). We apply it only on the current pixel and its neighbors while following a *4-connected* neighborhood scheme.

Post-treatment
In order to improve the detected region, any obtained agglomeration (non-significant small regions, holes and parasite pixels) that are located entirely within the region of interest are filled and aggregated to the pixels of the region.

Edge Detector
As edge detector, we have focused our choice on "Active contour" method. This choice is based on the fact that Active contours are closed and one-pixel thickness (i.e., they do not require post-treatments). However, we can remedy it major problem (not detecting of concave shapes) in the phase of Game theory integration.

In our implementation, thresholding image generated by the Otsu method [20] constitutes the input of the edge detector module. This Active contour is designed to fit the region in which the initial seed belongs.

Game Theory Integration
After running the two detectors for a sufficient number of iterations, Game theory can take place in order to improve the results of the two detectors cooperatively.

The involved pixels are those generated through the cooperation by results fusion between the edge and region detectors posing a decision-making problem. Thus, we first address the list of pixels located inside the Active contour and which does not belong to the region of interest (considering first the nearest pixel to the region of interest) (see Fig. 2). At the end of each iteration, an update of the region and the edge configurations is made.

Adding or removing a pixel to the region of interest is highly depending of the improvement or deterioration of the function value defined in formula (4).

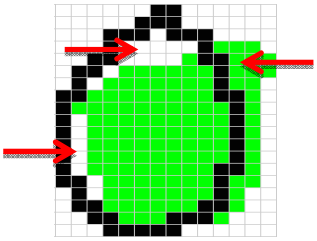


Fig. 1. Preliminary results and places where Game theory will be applied

4 Experimental Results

In this section, a summary of the tests and the obtained results is presented to demonstrate the effectiveness of our region-edge cooperation approach, we test it on a variety of different-kinds images (synthetic, real and medical: sane and added noise) from

two known images databases ([21], [22] and a set of MRI-type medical images found on the Internet). Also, we compare the individual approach results (Region growing only) to those of the proposed cooperative approach.

To evaluate the segmentation results, We will use the following methodology. First, we start our tests on the proposed approach by fixing a few parameters and varying the other, in a guided and judicious manner. This allows us to adjust the parameters of the various modules and study their impact on segmentation quality. Then, we test our cooperative approach to all the images of the three benchmarks, while determining for each image the region of interest to extract.

Obtained results are therefor compared and evaluated using the following criterion:

- **Borsotti criterion:** Uniformity and contrast. 0 nearest value represents best result.
- **Zeboudj criterion:** Contrast intra-inter region. 1 nearest value represents best result.

The comparison is done using the criteria mean values.

After testing our cooperative approach to sane images (net), and to discuss its robustness to noise, we propose to evaluate the same image after adding artificial noise.

Fig. 3 shows the results of the individual methods implementation (Active contour and Region growing) and the proposed cooperative approach on a real image from the BCU database [22]. Visual analysis of this figure shows that the cooperative segmentation using Game theory improves in parallel manner the results therefor obtained by correcting lacunas generated by the individual methods, namely the poor detection of concave regions in the Active contours and excess pixels presented in the result of segmentation by Region growing technique.

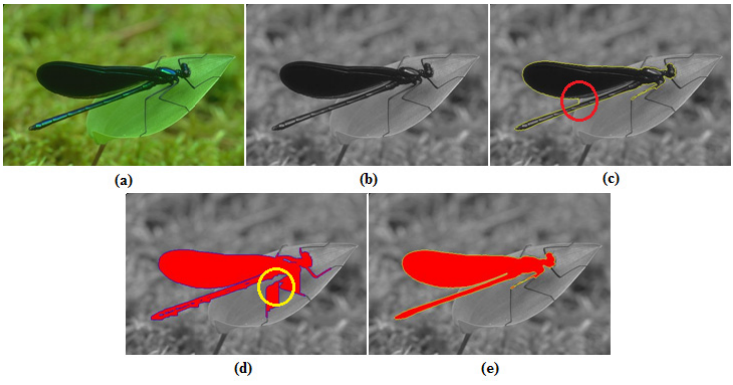


Fig. 2. Segmentation results of individual and cooperative approaches applied to a real image issued from the BCU database [22] (a) Original image (b) Grayscale image (c) Active Contour (d) Region growing (e) Image segmentation by region-edge cooperation using Game theory

4.1 Tests Results on Sane Images

The analysis of the registered segmentations results in terms of *Borsotti* and *Zeboudj* global mean values, allows us to go out with the following consequences:

1. As shown in the histogram of Fig. 4, the mean value of *Borsotti* criterion remains good and more or less stable for the set of all images used in the tests in the case of individual segmentation (Region growing) and the case of the integration of Game theory in the image cooperative segmentation.
2. Whereas, we observe an improvement in *Zeboudj* criterion in the second compared to the first which shows the effectiveness of the approach and the contribution of Game theory in image segmentation field.

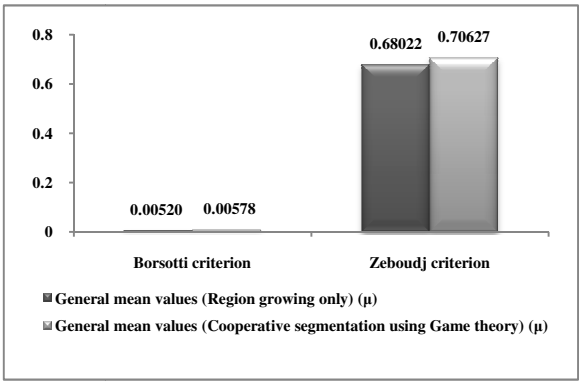


Fig. 3. General means values of Borsotti and Zeboudj criteria (without and with) integration of Game theory

4.2 Tests Results on Noisy Images

In this section we perform the operation of adding noise to three test images of different natures (Image (a) (synthetic), Image (b) (real) and Image (c) (medical)) while varying the percentage of added noise from 5% to 25%, which is randomly distributed over the whole of each of these images. Knowing that a percentage of 25% means that half of the image corresponds to noise (25% of pepper type and 25% of salt type); this represents a high rate of parasites pixels. The results obtained are illustrated in Fig.5.

Qualitative visual comparison between the original image and the segmented images produced by this technique shows that:

1. The degree of robustness to noise differs from one image to another; the synthetic and medical images have both a good robustness against noise varying from 5% to 25%, in which the quality of regions of interest segmentation is inversely proportional to the percentage of noise.
2. The real image segmentation result is good for percentages of 5% and 10% of added noise. However, it becomes very bad (invalid) from 15% of added noise.
3. Areas affected by the deterioration are often the borders rather than the interior of the region of interest (it is clearly visualized in the images (b) and (c)).
4. Generally, the qualities of the segmentation results provided by our approach applied to all three test images are reliable for a percentage of noise strictly less to 15%. This proves the robustness of the implemented procedure to the added noise.

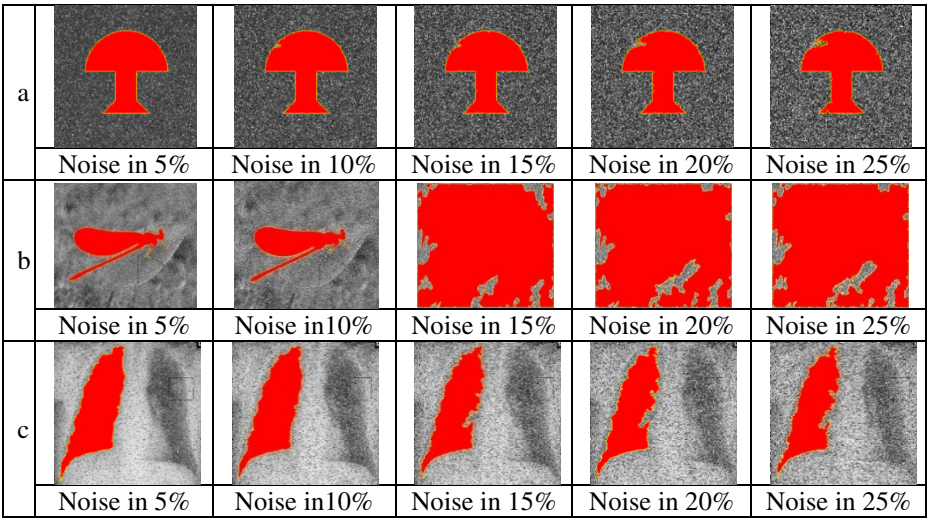


Fig. 4. Result of segmentation using a Game theory-based approach applied to different images after adding the salt and pepper noise (a) Synthetic image, (b) Real image, (c) Medical image

5 Conclusion and Discussions

In this article, we studied the possibility of Game theory integrating in image segmentation by region-edge cooperation. Indeed, we proposed a modified and simplified version of the parallel decision-making procedure as described in the work of Chacraborty and Duncan [1]. Whereas, the proposed modification helps to make the game cooperative, so that both players try coalitionary to improve the value of a common characteristic function within a framework of segmentation by results fusion.

Provided performance indices either digital or visual showed its effectiveness and its robustness to poor conditions of the input image (specifically image noise problem). Nevertheless, our method has had some inconveniences to running as:

- Results are depending on optimizing of parameters number, which is relatively big.
- Calculation time is sometimes very high estimated at a few hours.
- Procedure and by its nature can detect a single region of interest at a time.

Many prospects may be cited, for any enrichment of our study. Among them:

- Improvements of the detection procedure for all the regions of the image.
- Proposing of other gaming models, such as the players are pixels or image objects.
- Proposing image segmentation by mutual cooperation.
- Find an automatic parameters adjustment to the input image characteristics.

References

1. Chakraborty, A., Duncan, J.S.: Game-theoretic integration for image segmentation. IEEE Transaction on Pattern Analysis and Machine Intelligence, 12–30 (1999)
2. Bonnin, P., Zavidovique, B.: La segmentation coopérative: comment combiner détection de contours et croissance de régions? In: 14th Symposium Gresti Juan Les Pins (1993)
3. Haralick, R.M., Shapiro, L.G.: Image Segmentation techniques. Computer Vision Graphics Image Processing 29, 100–132 (1985)
4. Abdou, I.E., Pratt, W.K.: Quantitative design and evaluation of enhancement/thresholding edge detectors. Proceedings of the IEEE 67(5), 753–763 (1979)
5. Semchedine, M., Toumi, L.: Système Coopératif de Classification Floue Possibiliste avec Rejet d'Ambiguïté: Application à la segmentation d'images IRM. In: International Conference on Computer Integrated Manufacturing CIP (2007)
6. Acharya, T.A.: Image Processing, Principles and Applications: chapitre 7. A Wiley-Interscience Publication (2005)
7. Kornpobst, P.: Segmentation de régions, Odyssée Project. INRIA (1996)
8. Meliani, M.: Segmentation d'Image par Coopération Régions-Contours. Schoolmaster memory at National School of Computer Sciences, Algiers (2012)
9. Baillie, J.C.: Cours de Segmentation Module D9: traitement d'images et vision artificielle. ENSA (2003)
10. Maître, H.: Le traitement des images. Hermès, Traité IC2, Paris, France (2003)
11. Sebari, I., Dong-Chen, H.: Les approches de segmentation d'image par coopération région-contour. Revue Télédétection 7(1-2-3-4), 499–506 (2007)
12. Techno-Science.net.: Théorie des jeux : définition et explications, http://www.techno-science.net/?onglet=glossaire&definition=6426 (consulted on July 28, 2012)
13. Chaib-draa, B.: Chapitre 1 : Introduction à la Théorie des Jeux. Laval University, Computer Science & Software Engineering (CSSE) Department, Canada (2008), http://www.damas.ift.ulaval.ca/~coursMAS/ComplementsH10/Intro-TJ.pdf
14. Techno-Science.net.: Équilibre de Nash: définition et explications , http://www.techno-science.net/?onglet=glossaire&definition=6491
15. (consulted on July 28, 2012)
16. Chakraborty, A., Duncan, J.S.: Integration of boundary finding and region-based segmentation using game theory. In: XIVth International Conference on Information Processing in Medical Imaging, pp. 189–200 (1995)
17. Cassell, E., Kolar, S., Yakushev, A.: Using Game Theory for Image Segmentation (2007), http://www.angelfire.com/electronic2/cacho/machine-vision/ImSeg.pdf
18. Roy, K., Suen Ching, Y., Bhattacharya, P.: Segmentation of Unideal Iris Images Using Game Theory. In: ICPR 2010, pp. 2844-2847 (2010)
19. Ibragimov, B., Vrtovec, T., Likar, B., Pernus. F.: Segmentation of lung fields by game theory and dynamic programming. In: 4th International Workshop on Pulmonary Image Analysis - PIA 2011, Toronto, ON, Canada, September 18, pp. 101–111 (2011)
20. Kallel, M., Aboulaich, R., Habbal, A., Moakher, M.: A Nash-game approach to joint image restoration and segmentation. Applied Mathematical Modelling (2013), http://hal.inria.fr/hal-00648708
21. Otsu, N.: A threshold selection method from grey-level histograms. IEEE Transactions on Systems, Man, and Cybernetics SMC-9(1), 62–66 (1979)
22. BSDS500. Real images database found in, http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/BSR/BSR_bsds500.tgz (consulted on May 26, 2014)
23. Synthetic, B.: images database found in, http://pages.upf.pf/Sebastien.Chabrier/download/ImSynth.zip (consulted on May 27, 2014)

Improved Parameters Updating Algorithm for the Detection of Moving Objects

Brahim Farou[1,2(✉)], Hamid Seridi[2], and Herman Akdag[3]

[1] Computer Science Department, Badji Mokhtar-Annaba University, P.O.B 12, 23000 Annaba, Algeria
farou@ymail.com
[2] LabSTIC, Guelma University, POB 401, 24000 Guelma, Algeria
seridihamid@yahoo.fr
[3] LIASD, Paris 8 University, 93526 Saint-Denis, France
Herman.akdag@ai.univ-paris8.fr

Abstract. The presence of dynamic scene is a challenging problem in video surveillance systems tasks. Mixture of Gaussian (MOG) is the most appropriate method to model dynamic background. However, local variations and the instant variations in the brightness decrease the performance of the later. We present in this paper a novel and efficient method that will significantly reduce MOG drawbacks by an improved parameters updating algorithm. Starting from a normalization step, we divide each extracted frame into several blocks. Then, we apply an improved updating algorithm for each block to control local variation. When a significant environment changes are detected in one or more blocs, the parameters of MOG assigned to these blocks are updated and the parameters of the rest remain the same. Experimental results demonstrate that the proposed approach is effective and efficient compared with state-of-the-art background subtraction methods.

Keywords: Background subtraction · Motion detection · MOG · Machine vision · Videosurveillance

1 Introduction

The detection of moving object is the key step in many computer vision applications such as video surveillance, control applications, human machine interaction, and motion analysis. The challenge in such systems is to achieve high sensitivity in the detection of moving objects while maintaining a good discrimination rates and low processing time. The intrinsic nature of environment with illumination changes, shadows, waving flags, dust, bootstrapping and ghosts make tasks even more difficult. Recently, important efforts in this field have been focused on developing theories, methods and systems to deal with this problems and the most widely adopted techniques for handling these issues are optical flow, frame differencing and background subtraction. Background subtraction process is usually used with the assumption that the im-ages extracted form

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 527–537, 2015.
DOI: 10.1007/978-3-319-19578-0_43

video without any additional objects follow a fixed behavior and can be well described by a statistical model. In this case, the appearance of a new object in background will make this part inappropriate with the building model. The main idea in such approach is to model each pixel separately by a probability density function. Works done in [31] showed that GMM provides a good compromise between quality and execution time compared to other methods. The first use of GMM for modeling the background was proposed by Friedman and Russell [11].However Stauffer and Grimson [26] proposed the standard algorithm with an efficient update equations. Some extensions are given by [20,12,14] to improve the model adaptation speed. Other GMM algorithms were also proposed [27,34] to remove GMM drawback. Unfortunately, local variations and instant changes in brightness remains the major problem of GMM [33,13]. In the last decade, several studies have attempted to improve the performance of GMM in environments with multiple dimming and high condensation background. Initial ideas focused on substitution of using color characteristics [2] Setiawan et al. [24]or infrared camera [23]. Hybrid models such as GMM and K-means [3], GMM and fuzzy logic [1], Markov Random Fields [22],GMM and adaptive background [9,25], have been proposed to overcome GMM drawbacks. Other works have focused on improving the learning speed [15,28] through an adaptive learning rate [29], Better settings White and Shah [32] and the execution time [17] by using real parallel operations on multi-processor machines. Other systems use two backgrounds [4] to solve the problem of change in brightness between day and night or use Multi-level approaches [5,6,7]. Despite many algorithms have been pro-posed in the literature, the detection of moving objects in complex and dynamic environments is still far from being completely solved. In this paper, we will focus on the detection of moving objects in video surveillance through a fixed camera. To overcome the problems mentioned before, we propose a new and efficient background subtraction method based on GMM and local background monitoring. To cover all sections, the rest of the paper is organized as follows. The Preprocessing task is presented in Section 2. Section 3 is devoted to the similarity measurement. The background subtraction method and the proposed algorithm is presented in section 4 and section 6. We present in section 5 a local monitoring method used to update the MOG parameters. Results and discussion are presented in Section 6. Section 7 concludes the paper.

2 Preprocessing

The objective of the preprocessing task is to make the images more appropriate to apply algorithms in any system component allowing improvement in the success rates. In this phase, we start by transforming the captured video into a set of images. Then, we apply median filter to remove noise from the image. The extracted images from the video is done in the RGB color space, but this representation is not adequate because of the influence of light on the description of objects [28]. For this reason, we made a transfer to HSL model recognized to be one of the closest model of human perception and it provides a direct control of

chromaticity. The following are supplementary preprocessing techniques applied in our system.

2.1 Histogram Equalization

Histogram Equalization is an illumination normalization technique that uses the distribution of the original image to generate an image with uniform histogram. The objective of histogram equalization is to minimize the contrast in areas that are too light or too dark for an image.

2.2 Contour Detection

Contour correspond to the local variations in the intensity of the image pixel values. It is applied to preserve local features despite the influence of brightness. There are numerous contour detection techniques, but the context of real time processing lead to use a fast contour detection algorithm with inherent smoothing properties that can be adapted to different conditions of noise and artifacts. We used Soblel filter reported to be the best filter under real time consideration.

2.3 Splitting

This operation is only used in the initialization step. We divide the first frame into N equal size blocks to minimize local variations and to simplify the monitoring task. We noticed that the number of areas greatly influences on system quality. A large number of areas lead us to the starting point (pixel-based approach). In case where the number of areas is small (the size of the area is large), local variations accumulated in the same area force the system to consider the latter as an intense variation. In this way, all pixels belonging to the area will be updated. However, the number of blocks may change in processing time to improve system performance.

3 Similarity Measurement

The similarity between two sequences of measurement is a measure that quantifies the dependency between them. The use of similarity measure requires solving three major problems. The first one is to find the saved image that best matches the observed image. The second problem involves locating an object of interest in an observed image. The last one is the presence of rotational and scaling differences between the stored and observed image. In our case, the two first problems are similar and resolved by using contour detection algorithm. Indeed, the original image is divided into a set of blocks and the similarity is applied, not to detect any type of object, but to measure the blocks dependence at the same position between the reference image and the observed image. The use of a binary image containing only contours, reduces the brightness change effect since the contours are invariant to the latter. The third problem is not probable since

the camera is static and it has no zoom effect. Various similarity measures have been proposed in the literature. However, each measure has its own strengths and weaknesses and a measure that performs well on one type of images may not work on another types. In this paper we use Pearsons correlation coefficient which is reported in the literature as the best similarity measure on various image types.

4 Mixture of Gaussians

MoG is a statistical model that assumes the data where originates from a weighted sum of several Gaussian distributions. Stauffer and Grimson [26] presented an adaptive GMM method to model a dynamic background in image sequences. If K Gaussian distributions are used to describe the history of a pixel, the observation of the given pixel will be in one of the K states at one time [3]. K determines the multimodality of the background and the selection of K is generally based on the available memory and computing power. Stauffer and Grimson [26] proposed to set K from 3 to 5. First, each pixel is characterized by its intensity in the HSL color space. Then, the probability of observing the current pixel value is given by the following equation in the multidimensional case:

$$P\left(P_t\right) = \sum_{i=1}^{k} w_{i,t} \cdot \eta\left(P_t, \mu_{i,t}, \Sigma_{i,t}\right) \tag{1}$$

Where: k is the number of associated Gaussians to each pixel, $w_{i,t}$ is the calculated weight, $\mu_{i,t}$ is the mean and $\Sigma_{i,t}$ is the covariance matrix that are respectively evaluated for the ith Gaussian at time t. η is a Gaussian probability density function:

$$\eta\left(P_t, \mu, \Sigma\right) = \frac{1}{2\pi^{\frac{n}{2}} \left|\Sigma\right|^{\frac{1}{2}}} \exp^{\frac{1}{2}\left(P_t-\mu\right)\Sigma^{-1}\left(P_t-\mu\right)} \tag{2}$$

For real time consideration, the update of the model is carried out by using an online K-Means approximation algorithm [3], [8]. After the parameters initialization, a first foreground detection can be made and the parameters are updated. When the new frame incomes, each pixel value is checked through the existing k Gaussian distributions, until a match is found. A pixel matched a Gaussian distribution if the pixel value is within 2.5 standard deviations of distribution according to Eq. 3.

$$\frac{\left|P_t - \mu_i\right|}{\sigma_i} < 2.5 \tag{3}$$

When a match is found with one of the k Gaussian, we look for the Gaussian distribution classification. If the Gaussian distribution is identified as a background, the pixel is classified as background. Otherwise, the pixel is classified as

foreground. The prior weights of the K distributions are updated according to Eq. 4:

$$W_{k,t} = (1 - \alpha) \cdot W_{k,t-1} + \alpha M_{k,t} \tag{4}$$

Where: α is the learning coefficient which determines the model adaptation speed and $M_{k,t}$ is equal to 1 for the distribution which satisfy 3and 0 for others. After updating weights, a normalization step is carried out to ensure that the sum of the weights is always equal to 1. For the unmatched components, μ and σ parameters remain unchanged. The parameters of the distribution which matches the new observation are updated using the following equations:

$$\mu_{k,t} = (1 - \varphi_k) \cdot \mu_{k,t-1} + \varphi_k \cdot P_t \tag{5}$$

$$\sigma^2_{k,t} = (1 - \varphi_k) \cdot \varphi^2_{k,t-1} + \varphi_k \left(P_t - \mu_{k,t}\right)^T \left(P_t - \mu_{k,t}\right) \tag{6}$$

With

$$\varphi_t = \alpha\eta \left(P_t / \mu_k \sigma_k\right) \tag{7}$$

If none of the distributions satisfy the Eq. 3, then the pixel is associated with first plan and the parameters of the least probable distribution is replaced by a new Gaussian with the current value as its mean value, an initially high variance, and a low prior weight parameter according to Eq. 8,Eq. 9 and Eq. 10 described below:

$$\sigma^2_{k,t} = Large \quad Initial \quad Variance \tag{8}$$

$$W_{k,t} = Low \quad Prior \quad Weight \tag{9}$$

$$\mu_{k,t} = P_t \tag{10}$$

W is the initial weight value for the new Gaussian. If w is higher, the distribution is chosen as the background model for a long time. To decide if P_t is included in the background distributions, the distributions are ordered according to the value of $W_{k,t}/\sigma_{k,t}$. This ordering use the assumption that a background pixel corresponds to a high weight with a weak variance for the reason that the background is practically constant and it is more present than moving objects. The first β distributions that verify the Eq. 11 are selected to represent the background.

$$\beta = \arg\min \left(\Sigma_{k=1}^{b} W_{k,t} > B\right) \tag{11}$$

The threshold B represents the minimum portion of the total weight given to background model. If a small value for B is chosen, then the background becomes unimodal. If B is higher, a multi-modal distribution caused by a repetitive background motion could result from a variety of background component that allows the background to accept more than one Gaussian distribution. The use of unique threshold B for GMM implies a miss classification especially when scene contains both dynamic and static area. A higher threshold can achieve correct classifications in a dynamic background but makes incorrect detection of moving objects in stationary background.

5 Adaptive Local Monitoring

Methods based on MOG use the pixel value for detecting a probable change in the background based on the assumption that a moving object is a set of pixels in movement. This vision is very useful because it requires no a priori knowledge of objects and their trajectories. However, the natural environment is far from perfect. The presence of dust, the change in brightness, rain, wind,etc. influence on pixel value making unwanted local variation and leading to a misdetection of motion. The false pixel detection induces the system to make errors in the following steps, either by the deformation of the moving objects or by signaling a false movement. To overcome these problems, we proposed an adaptive local monitoring algorithm for each block to control local variation. From the start of the process of detecting moving objects, the monitoring task is enabled by assigning an observer to each block. The role of the latter is to monitor and report the presence of any activity that may be a movement. The decision of the presence or absence of movement is ensured by calculating the similarity between two states of the same block. Indeed, after assigning a block for each monitor, it stores the initial state which contains only contour. The first state is taken without the presence of moving objects. The second state represents the image in process. For the convenience of the update algorithm, each pixel has been labeled with the block number to which it belongs. It is used to provide updates to the concerned pixels only. So if a significant activity in a block of the image is detected, the parameters of the Gaussian assigned to all pixels of this block decide whether there has been any motion, and will be updated according to the proposed model. The parameters of the Gaussian assigned to other blocks will not undergo any change. This process will eliminate local variation, because only blocks with significant change will be considered by the system.

6 Results and interpretation

The system presented in this paper is implemented in Java on a computer with an Intel Core i5 2.67 GHz and a 4GB memory capacity. In this section, we shall present results of our method while challenging real-world situations. We take in addition to our database, three publicly available Benchmark Dataset Collection. The first one (BDC1) has six sequences in the Dataset (campus, highway I, highway I2, highway II intelligent room and laboratory) [21]. The second (BDC2) has nine sequences (bootstrap, a campus, a curtain, an escalator, a fountain, a hall, a lobby, a shopping mall and a water surface) [18].The last one (BDC3) has two sequences (highway and hallway) [16]. Our database (BDC4) has four sequences (campus, a hallway, a highway and a public park). In BDC4, The outdoor videos are recorded in a random situation and without any assumption on the observed scene where a group of clouds is passing in the sky, causing sudden illumination changes. For measuring accuracy we used different metrics, namely Precision and Recall.

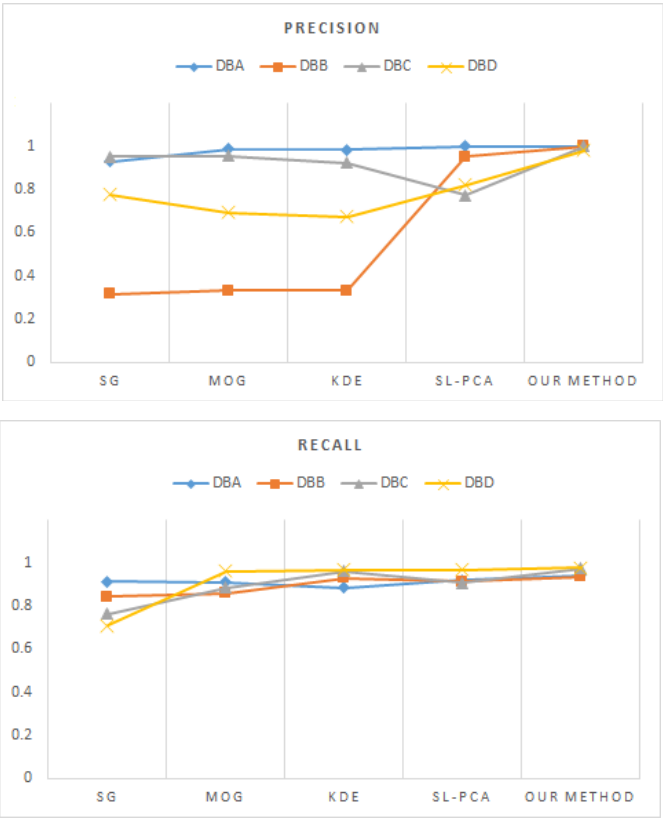


Fig. 1. Precision and recall results for MOG [26], SL-PCA [19], SG [30], KDE [10], and our method

Fig. 2. Background subtraction results in personnel video database in both indoor and outdoor environments

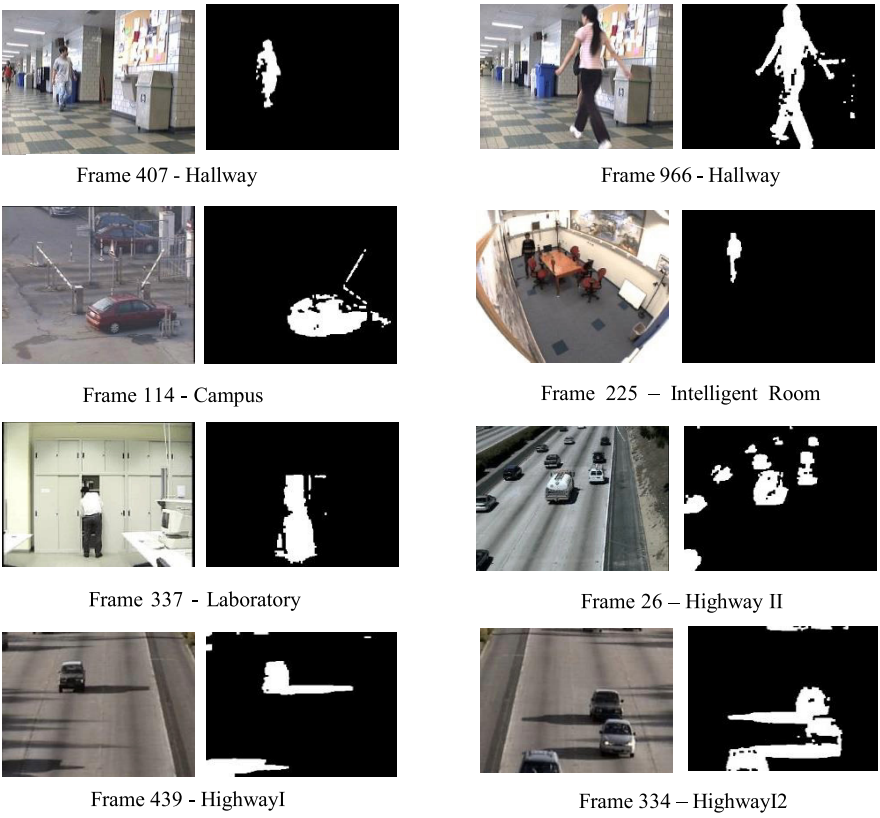


Fig. 3. Background subtraction results in public video database

Recall gives the percentage of corrected pixels classified as background when compared with the total number of background pixels in the ground truth. Precision gives the percentage of corrected pixels classified as background as compared at the total pixels classified as background by the method.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

FP and FN refer to pixels misclassified as foreground (FP) or background (FN) while TP account for accurately classified pixels respectively as foreground. A good performance is obtained when the detection rate is high without altering the precision.

Figure.1 clearly shows that our method outperform the well-known background subtraction methods in term of precision and recall. Figure.2 and Figure.3 show some qualitative results on both public and personal databases. They

show that our system is able to give a very good subtraction in environment without any assumption on lighting condition. They also show the effectiveness of the proposed system in outdoor and indoor environment.

7 Conclusion

In this paper, we proposed a background subtraction system for image sequences extracted from fixed camera using an improved parameters updating algorithm for mixture of Gaussians. To overcome the brightness and local variation we first made a transition from RGB to HSL color space. Then we divided the image into N areas and assigned to each one a local monitoring algorithm that allows selecting regions with a very large change using Pearsons Correlation Coefficient. Transfer to HSL color space has significantly decreased light effect on the system behavior through accumulating all brightness variations in a single component (L). While segmenting the image into regions have eliminated local variations caused mainly by the presence of dust. Tests conducted on databases show that our system has a good sensitivity, more accuracy compared with well-known methods. In future work, our algorithm will be adjusted by dividing the image into homogenous regions and solving the problem of shadow and color similarity between moving objects and background.

References

1. Baf, F.E., Bouwmans, T., Vachon, B.: Fuzzy statistical modeling of dynamic backgrounds for moving object detection in infrared videos. In: Computer Vision and Pattern Recognition (2009)
2. Caseiro, R., Henriques, J.F., Batista, J.: Foreground Segmentation via Background Modeling on Riemannian Manifolds. In: International Conference on Pattern Recognition, pp. 3570–3574 (2010)
3. Charoenpong, T., Supasuteekul, A., Nuthong, C.: Adaptive background modeling from an image sequence by using K-Means clustering (2010)
4. Cheng, F.C., Huang, S.C., Ruan, S.J.: Illumination-Sensitive Background Modeling Approach for Accurate Moving Object Detection. IEEE Transactions on Broadcasting 57, 794–801 (2011)
5. Cristani, M., Bicegi, M., Murino, V.: Integrated Region-and Pixel-based Approach to Background Modeling (2002)
6. Cristani, M., Murino, V.: A spatial sampling mechanism for effective background subtraction. In: Computer Vision Theory and Applications, pp. 403–412 (2007)
7. Cristani, M., Murino, V.: Background Subtraction with Adaptive Spatio-Temporal Neighborhood Analysis. In: Computer Vision Theory and Applications, pp. 484–489 (2008)
8. Djouadi, A., Snorrason, G.F.D.: The Quality of Training Sample Estimates of the Bhattacharyya Coefficient. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 92–97 (1990)
9. Doulamis, A., Kalisperakis, I., Stentoumis, C., Matsatsinis, N.: Self Adaptive background modeling for identifying persons' falls. In: International Workshop on Semantic Media Adaptation and Personalization (2010)

10. Elgammal, A.M., Harwood, D., Davis, L.S.: Non-parametric Model for Background Subtraction (2000)
11. Friedman, N., Russell, S.J.: Image Segmentation in Video Sequences: A Probabilistic Approach. In: Uncertainty in Artificial Intelligence, pp. 175–181 (1997)
12. Hayman, E., Olof Eklundh, J.: Statistical Background Subtraction for a Mobile Observer. In: International Conference on Computer Vision, pp. 67–74 (2003)
13. Hedayati, M., Zaki, W.M.D.W., Hussain, A.: Real-time background subtraction for video surveillance: From research to reality. In: International Colloquium on Signal Processing & Its Applications (2010)
14. Kaewtrakulpong, P., Bowden, R.: An improved adaptive background mixture model for real time tracking with shadow dectection (2001)
15. Kan, J., Li, K., Tang, J., Du, X.: Background modeling method based on improved multi-Gaussian distribution. In: International Conference on Computer Application and System Modeling (2010)
16. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Statistical modeling of complex backgrounds for foreground object detection. IEEE Transactions on Image Processing 13, 1459–1472 (2004)
17. Li, X., Jing, X.: FPGA based mixture Gaussian background modeling and motion detection 4, 2078–2081 (2011)
18. Martel-brisson, N., Zaccarin, A.: Learning and Removing Cast Shadows through a Multidistribution Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1133–1146 (2007)
19. Oliver, N.M., Rosario, B., Pentland, A.P.: A Bayesian Computer Vision System for Modeling Human Interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 831–843 (2000)
20. Power, P.W., Schoonees, J.A.: Understanding Background Mixture Models for Foreground Segmentation (2002)
21. Prati, A.: c, I.M., Trivedi, M.M., Cucchiara, R.: Detecting Moving Shadows: Formulation, Algorithms and Evaluation
22. Schindler, K., Wang, H.: Smooth Foreground-Background Segmentation for Video Processing (2006)
23. Seki, M., Okuda, H., Hashimoto, M., Hirata, N.: Object modeling using gaussian mixture model for infrared image and its application to vehicle detection. Journal of Robotics and Mechatronics 18(6), 738 (2006)
24. Setiawan, N.A., Ju Hong, S., Woon Kim, J., Woo Lee, C.: Gaussian Mixture Model in Improved HLS Color Space for Human Silhouette Extraction (2006)
25. Sheng, Z.B., Cui, X.Y.: An adaptive learning rate GMM for background extraction. Optoelectronics Letters 4, 460–463 (2008)
26. Stauffer, C., Grimson, W.E.L.: Adaptive Background Mixture Models for Real-Time Tracking. Computer Vision and Pattern Recognition 2, 2246–2252 (1999)
27. Stenger, B., Ramesh, V., Paragios, N., Coetzee, F., Buhmann, J.M.: Topology Free Hidden Markov Models: Application to Background Modeling. In: International Conference on Computer Vision, pp. 294–301 (2001)
28. Suo, P., Wang, Y.: An improved adaptive background modeling algorithm based on Gaussian Mixture Model. In: International Conference on Signal Processing Proceedings (2008)
29. Wang, H., Suter, D.: A re-evaluation of mixture of Gaussian background modeling [video signal processing applications]. In: International Conference on Acoustics, Speech, and Signal Processing, vol. 2 (2005)

30. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-Time Tracking of the Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 780–785 (1997)
31. Yu, J., Zhou, X., Qian, F.: Object kinematic model: A novel approach of adaptive background mixture models for video segmentation (2010)
32. Zang, Q., Klette, R.: Evaluation of an Adaptive Composite Gaussian Model in Video Surveillance (2003)
33. Zhang, L., Liang, Y.: Motion Human Detection Based on Background Subtraction. In: International Workshop on Education Technology and Computer Science (2010)
34. Zivkovic, Z., Heijden, F.V.D.: Recursive Unsupervised Learning of Finite Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 651–656 (2004)

Towards Real-Time Co-authoring of Linked-Data on the Web

Moulay Driss Mechaoui(✉), Nadir Guetmi, and Abdessamad Imine

[1] University of Sciences and Technology Oran 'Mohamed Boudiaf' USTO-MB
Mathematics and Computer Science Faculty
Oran, Algeria
[2] LIAS/ISAE-ENSMA, Poitiers University
Chasseneuil, France
[3] Université de Lorraine and INRIA-LORIA Grand Est
Nancy, France
moulaydriss.mechaoui@univ-usto.dz,
nadir.guetmi@ensma.fr,
abdessamad.imine@loria.fr

Abstract. Real-time co-authoring of Linked-Data (LD) on the Web is becoming a challenging problem in the Semantic Web area. LD consists of RDF (Resource Description Framework) graphs. We propose to apply state-of-the art collaborative editing techniques to manage shared RDF graphs and to control the concurrent modifications. In this paper, we present two concurrency control techniques. The first one is based on client-server architecture. The second one is more flexible as it enables the collaborative co-authoring to be deployed in mobile and P2P architecture and it supports dynamic groups where users can leave and join at any time.

Keywords: Linked-Data · Collaborative editing systems · Optimistic replication

1 Introduction

Recently, providing collaborative co-authoring tools in the Web Semantic is becoming more attractive as they enable semantic web data to be produced in online mode and to be available to a large public. Linked Data (LD) is recently used to replace collections of offline RDF data [3]. The goal of LD is to enable people to share structured data on the web as easily as they can share documents today. It uses RDF technology that (i) relies on HTTP URIs to denote things; (ii) provides useful information about a thing at that thing's URI; and (iii) includes in that information other URIs of LD. Tabulator [2] is a LD browser, designed to provide the ability to navigate the web of linked things. In [3], Berners-Lee et al. raise some interesting challenges when adding collaborative co-authoring mode in Tabulator. This mode consists in collaboratively editing the LD which is represented by a RDF graph.

In this paper, we sketch two solutions that may meet to some extent the read-write requirement in LD browser. We consider a RDF graph as a shared data

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 538–548, 2015.
DOI: 10.1007/978-3-319-19578-0_44

which can be edited and updated by several users. To control the concurrent access to this shared data, we propose to apply state-of-the art collaborative editing techniques [9,6]. The CRDT (Commutative Replicated Data Type) is a class of algorithms that is emerging for ensuring consistency of highly dynamic content on P2P networks. However, this approach incurs some overhead they do not consider directly a set as a list (or a sequence) [1]. Also, with the continuously growing amount of structured data available on the Semantic Web there is an increasing desire to replicate such data to mobile devices. This enables services and applications to operate independently of the network [18,11]. Classical replication techniques cannot be properly applied to mobile systems because they do not adopt to changing user information needs, and they do not consider the technical, environmental, and infrastructural restrictions of mobile devices. We think that Operational Transformation (OT) approach [4,14] may be a good candidate as it supports unconstrained interaction. Indeed, it allows any user to modify any shared data consistently at any time without any restrictions on users's actions.

The rest of the paper is organized as follows. Section 2 presents the ingredients of OT approach. In Section 3, we suggest two concurrency control procedures for managing the collaborative edition of RDF graphs. Section 4 discusses performance evaluation, and concludes.

2 Transformational Approach

Principle. Operational Transformation (OT) is an optimistic replication technique which allows many users (or sites) to concurrently update the shared data and next to synchronize their divergent replicas in order to obtain the same data [17]. The updates of each site are executed on the local replica immediately without being blocked or delayed, and then are propagated to other sites to be executed again. Accordingly, every update is processed in four steps: (i) *generation* on one site; (ii) *broadcast* to other sites; (iii) *reception* on one site; (iv) *execution* on one site.

A crucial issue when designing shared data with a replicated architecture and arbitrary messages communication between sites is the *consistency maintenance* (or *convergence*) of all replicas. To illustrate this problem, consider the following example:

Example 1. Consider the following group text editor scenario (see Figure 1.(a)): there are two users (on two sites) working on a shared document represented by a sequence of characters. These characters are addressed from 0 to the end of the document. Initially, both copies hold the string "*efecte*". User 1 executes operation $op_1 = Ins(1, f)$ to insert the character f at position 1. Concurrently, user 2 performs $op_2 = Del(5)$ to delete the character e at position 5. When op_1 is received and executed on site 2, it produces the expected string "*effect*". But, when op_2 is received on site 1, it does not take into account that op_1 has been executed before it and it produces the string "*effece*". The result at site 1 is different from the result of site 2 and it apparently violates the intention of

op_2 since the last character e, which was intended to be deleted, is still present in the final string. Consequently, we obtain a *divergence* between sites 1 and 2. It should be pointed out that even if a serialization protocol [4] was used to require that all sites execute op_1 and op_2 in the same order (*i.e.* a global order on concurrent operations) to obtain an identical result *effece*, this identical result is still inconsistent with the original intention of op_2.

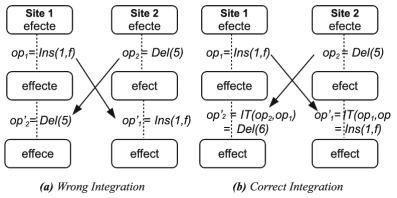


Fig. 1. Serialization of concurrent updates

To maintain convergence, the OT approach has been proposed by [4]. When User X gets an operation op that was previously executed by User Y on his replica of the shared object User X does not necessarily integrate op by executing it "as is" on his replica. He will rather execute a variant of op, denoted by op' (called a *transformation* of op) that *intuitively intends to achieve the same effect as op*. This approach is based on a transformation function (or algorithm) IT that apply to couples of concurrent operations defined on the same state.

Example 2. In Figure 1.(b), we illustrate the effect of IT on the previous example. When op_2 is received on site 1, op_2 needs to be transformed according to op_1 as follows: $op'_2 = IT((Del(5), Ins(1, f)) = Del(6)$. The deletion position of op_2 is incremented because op_1 has inserted a character at position 1, which is before the character deleted by op_2. Next, op'_2 is executed on site 1. In the same way, when op_1 is received on site 2, it is transformed as follows: $IT(Ins(1, f), Del(5)) = Ins(1, f)$; op_1 remains the same because f is inserted before the deletion position of op_2.

Intuitively we can write the transformation IT as follows:

```
IT(Ins(p1,c1),Ins(p2,c2)) =
          if (p1 < p2) return Ins(p1,c1)
          else return Ins(p1+1,c1)
          endif;
```

OT Model. Using the OT approach, each site is equipped by two main components [4,10]: the *integration component* and the *transformation component*. The integration component determines how an operation is transformed against a given operation sequence (*e.g.*, the log buffer). It is also responsible for receiving, broadcasting and executing operations. It is rather *independent* of the

type of the shared data. The transformation component is a set of IT algorithms which is responsible for merging two concurrent operations defined on the same state. Every IT algorithm is *specific* to the semantics of a given shared data.

The most known OT-based theoretical framework is established by Ressel et al. [10]. They define two consistency criteria:

- **Causality.** If one operation op_1 causally precedes another operation op_2, then op_1 must be executed before op_2 at all sites.
- **Convergence.** When all sites have performed the same set of operations, the copies of the shared data must be identical.

It has been proved that any integration component can achieve convergence in the presence of arbitrary transformation paths if its IT algorithm satisfies two properties $TP1$ and $TP2$ [10]. For all op, op_1 and op_2 pairwise concurrent operations with $op'_1 = IT(op_1, op_2)$ and $op'_2 = IT(op_2, op_1)$:

- **TP1**: $[op_1 ; op'_2] \equiv [op_2 ; op'_1]$.
- **TP2**: $IT(IT(op, op_1), op'_2) = IT(IT(op, op_2), op'_1)$.

Property $TP1$ defines a *state identity* and ensures that if op_1 and op_2 are concurrent, the effect of executing op_1 before op_2 is the same as executing op_2 before op_1. This property is necessary but not sufficient when the number of sites is greater than two. Property $TP2$ defines an *update identity* and ensures that transforming op along equivalent and different operation sequences will give the same operation.

Properties $TP1$ and $TP2$ are sufficient to ensure the convergence for *any number* of concurrent operations which can be executed in *arbitrary order* [10]. Accordingly, by these properties, it is not necessary to enforce a global total order between concurrent operations because data divergence can always be repaired by operational transformation. However, finding an IT algorithm that satisfies $TP1$ and $TP2$ is considered as a hard task, because this proof is often unmanageably complicated [13]. To overcome this difficulty, we proposed in [6] a formal methodology for designing and analyzing IT algorithms by using a theorem prover.

Several OT-based integration components have been proposed in the groupware research area. These components may be categorized in two categories. The first one does not require $TP2$ property: it relies on client-server architecture for enforcing a unique transformation order. We can cite in this category algorithms like SOCT4 [16] and TIBOT [7]. As for the second category, it requires $TP2$ property. This constraint enables the concurrent operations to be synchronized in a decentralized way. Algorithms such as adOPTed [10] SOCT2,4 [15,16] and GOTO [14] belong to this category.

3 Our Proposals

To manage all concurrent access for editing collaboratively a shared RDF graph, we need a concurrency control procedure. In this section, we first argue how to map a RDF graph into a sequence data structure. According to centralized and decentralized architectures, we suggest two concurrency control procedures.

3.1 RDF Graph as a Sequence

When publishing LD on web, information about resources is represented using the RDF. Any expression in RDF is a collection of *triples*, each consisting of a *subject*, a *predicate* (also called property) and an *object*. The subject of a triple is the URI describing resource. The object can either be a simple literal value (*e.g.*, a string, a number) or the URI of another resource. The predicate indicates what kind of relation exists between subject and object. The predicate is a URI too. A set of such triples is called an RDF graph. This can be illustrated by a node and directed-arc diagram, in which each triple is represented as a node-arc-node link.

Usually a set is implemented by means of a list. It means we can use operations, such as insert and delete, to edit a shared list. Thus, we can reuse the state-of-the-art of collaborative editing systems.

For instance, the following three english statements (this example is taken from [8]):

- <http://www.example.org/index.html>has a creator whose value is John Smith
- <http://www.example.org/index.html>has a creation-date whose value is August 16, 1999
- <http://www.example.org/index.html>has a language whose value is English

could be represented by the RDF graph shown in Figure 2.

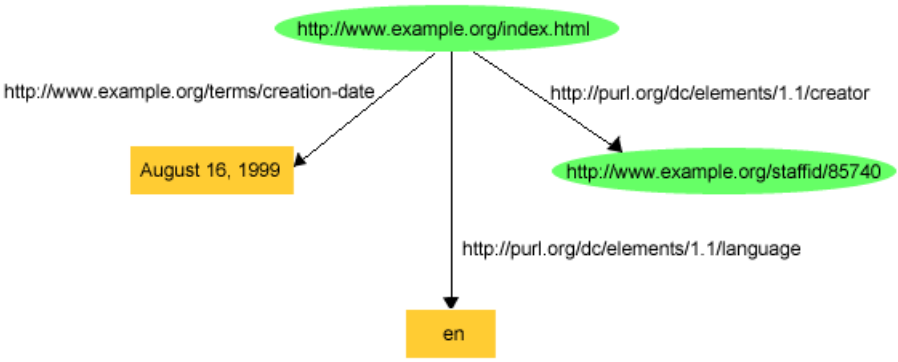


Fig. 2. An RDF Graph

An RDF graph can be serialized into a sequence of triples and considered as a text where each line corresponds to a simple triple of subject, predicate and object. For example, the third statement shown in Figure 2 would be written as a text line:

```
<http://www.example.org/index.html>
                <http://purl.org/dc/elements/1.1/language> "en" .
```

By considering an RDF graph as a sequence, each triple is addressed simply by a position within the sequence. Therefore, we assume that the sequence of triples can be modified by the following primitive operations:

- $Ins(p, t)$ which adds triple t at position p;
- $Del(p, t)$ which deletes triple t at position p.

Updating a triple (*e.g.*, by modifying the predicate URI) can be expressed by a sequence of delete (by removing the old triple) and insert (by adding the new one) operations.

3.2 Ingredients of Collaboration

Each user's site has a local copy of RDF graph and a unique identity. We assume that the RDF graph is serialized in the same way on every site.

Every site generates operations sequentially and stores these operations in a stack also called a *log*. When a site receives a remote operation *op*, the integration component executes the following steps:

1. from the local log it determines the sequence *seq* of operations that are concurrent to *op*;
2. it calls the transformation component in order to get operation op' that is the transformation of *op* according to *seq*;
3. it executes op' on the current state;
4. it adds op' to the local log.

3.3 Centralized Solution

In this section, we propose a real-time co-authoring based on client-server architecture. Indeed, users can edit collaboratively a shared RDF graph by reconciliating their divergent copies via a particular site called *server*. We think that SOCT4 [16] is most appropriate to this kind of architecture.

In SOCT4, the operations are ordered globally by using a timestamp given by the server. When an operation is generated on site s, it is immediately executed (to satisfy real-time constraint), but it is not propagated until it gets a timestamp from the server and all the operations which precede it according to the timestamp order have been received and executed on s. Moreover, this operation is transformed against all concurrent operations (operations received after its generation and preceding it in the global order) before to be propagated. To ensure convergence, SOCT4 requires only the property $TP1$ to be satisfied by the IT algorithm.

Example 3. Consider two users editing a shared RDF graph as described in Figure 3. Initially, each site has an empty copy. The index of each operation represents the timestamp given by the server. Two local insertion operations op_1 and op_2 have been executed by user 1 (at site 1). Concurrently, user 2 has executed another insertion operation op_3. The added triples t_1, t_2 and t_3 are

$site\,1$	$site\,2$
$op_1 = Ins(0, t_1)$	$op_3 = Ins(0, t_3)$
$op_2 = Ins(1, t_2)$	
$s_1 = synchronize$	
	$s_2 = synchronize$
$s_3 = synchronize$	

Fig. 3. Scenario of collaboration

respectively as follows (where `UR1` is `http://www.example.org/index.html` and `UR2` is `http://www.example.org/staffid/85740`):

```
<UR1> <http://www.example.org/terms/creation-date> "August 16, 1999" .

<UR1> <http://purl.org/dc/elements/1.1/language> "en" .

<UR1>  <http://purl.org/dc/elements/1.1/creator> <UR2> .
```

1. At point s_1, site 1 decides to synchronize with other sites. As there is no concurrent operation available, op_1, op_2 are sent to site 2 (via the server) in their original forms.

2. At point s_2, site 2 cannot send op_3 as long as it did not receive the precedent operations (according to the timestamp order). Thus the synchronization calls IT algorithm to produce the following transformations:

$op'_1 = IT(op_1, op_3) = Ins(0, t_1)$
$op'_3 = IT(op_3, op_1) = Ins(1, t_3)$
$op'_2 = IT(op_2, op'_3) = Ins(1, t_2)$
$op''_3 = T(op'_3, op_2) = Ins(2, t_3)$

op'_1, op'_2 are executed on site 2, and op''_3 is broadcast to other sites.

3. At point s_3, site 1 decides again to synchronize. The remote operation op''_3 is executed directly (without transformation) after op_1 and op_2.

4. Note that, after point s_3, sites 1 and 2 have the same log, namely op_1, op_2 and op''_3. However, site 1 has performed the following sequence:

op_1
op_2
$op''_3 = IT(IT(op_3, op_1), op_2)$

while site 2 has executed the following sequence:

op_3
$op''_1 = IT(op_1, op_3)$
$op''_2 = IT(op_2, op'_3)$

As SOCT4 requires only $TP1$ property, the above sequences are equivalent in the sense that they produce the same RDF graph. The operations are stored in the log according to the timestamp order but they may be executed in different orders at different sites.

It should be noted that SOCT4 has been used successfully in the development of a File Synchronizer [9] distributed with the industrial collaborative development environment, LibreSource Community[1], proposed by ARTENUM Company. LibreSource is a platform for hosting virtual teams. Users can register and create channels for synchronizing shared data. On a single server, LibreSource can host several projects, several groups of users, and grant fine grain access to the resources.

Although SOCT4 ensures causality and convergence properties, it degrades the responsiveness of the system as all messages are exchanged via a server. Moreover, it does not scale because it is based on a single point of failure.

3.4 Decentralized Solution

Integration algorithms based on $TP2$ property enable concurrent operations to be synchronized in a decentralized way. Thus, they avoid a single point of failure. Nevertheless, these algorithms have limited scalability with the number of users. Indeed, all proposed OT frameworks rely on a fixed number of users during collaboration sessions. This is due in the fact that they use vector timestamps to enforce causality dependency. The vector timestamps do not scale well, since each timestamp is a vector of integers with a number of entries equal to the number of users.

In [5], we proposed a new framework for collaborative editing to address the weakness of previous OT works. The features of our framework are as follows:

1. It supports an unconstrained collaborative editing work (without the necessity of central coordination). Using optimistic replication scheme, it provides simultaneous access to shared data.
2. Instead of vector timestamps, we use a simple technique to preserve causality dependency. Our technique is minimal because only direct dependency information between operations is used. It is independent on the number of users and it provides high concurrency in comparison with vector timestamps.
3. Using OT approach, reconciliation of divergent copies is done automatically in decentralized fashion.
4. Our framework can scale naturally thanks to our minimal causality dependency relation. In other words, it may be deployed easily in Peer-to-Peer (P2P) networks.

Example 4. Consider the scenario given in Example 3. In our framework, operations op_1 and op_2 will be related by a dependency. This is due in the fact that their added triples are adjacent (positions 0 and 1) and created by the same user. Thus, op_1 must be executed before op_2 at all sites. This dependency relation is minimal in the sense that when op_2 is broadcast to all sites it holds only the identity of op_1 as it depends on directly.

1. At site 1, op_3 is considered as concurrent. It is then transformed against op_1 and op_2. The following sequence is executed and logged in site 1:

[1] http://dev.libresource.org

$op_1 = Ins(0, t_1)$
$op_2 = Ins(1, t_2)$
$op_3'' = IT(IT(op_3, op_1), op_2) = Ins(2, t_3)$

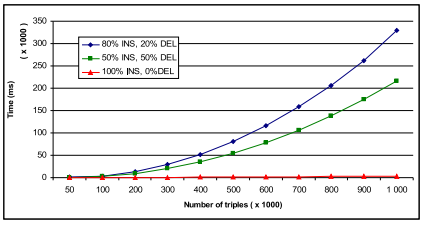
2. At site 2, op_1 and op_2 are concurrent with respect to op_3. They must be transformed before to be executed after op_3 according to their dependency relation. Thus, the following sequence is executed and logged in site 2:

$op_3 = Ins(0, t_3)$
$op_1' = IT(op_1, op_3) = op_1$
$op_2' = op_2$

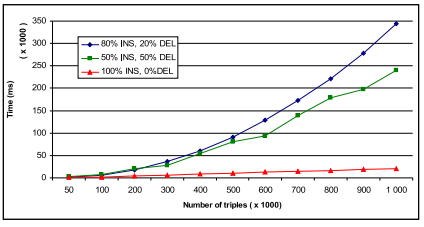
Unlike the others OT-based integration algorithms, we minimize the transformation steps when integrating a remote operation depending on another operation. Indeed, at site 2, the new form of op_2 is deduced from the executed form of op_1 (without transformation as in Example 3). On the other hand, the sequences of sites 1 and 2 are not identical but equivalent.

4 Performance Evaluation

Our experimentation consists to compare the response time of generating and integrating a sequence of remote triples over a local ones. We use two sites (Site1 and Site2), initially the log of each sites is empty. Each site generates locally a sequence of operations; the sites communicate the generated operations to be integrated. The sizes of the sequence are varied from 50 000 to 1 000 000 triples. The percentage of insertions in the sequence and the log are variants from 50%, 80% to 100%.



(a) Generation + Integration time of a sequence of triples over an empty RDF document



(b) Generation + Integration time of a sequence of triples over a RDF document containing 10 000 triples

Fig. 4. Updating RDF document

We implement a prototype of Optic [5] in java, compiled by NetBeans 6.8 with JVM heap size 1GB, and executed on a computer running Windows XP SP2 with an Intel (R) Core (TM) 2 CPU E7400 @ 2.80 GHz and 2 GB RAM. We calculate the sum of the generation time of the sequence in the Site1 with the

time of integration of the same sequence in the Site2. For every generation and integrating sequence three times are executed and the average time is recorded.

The Figure 4(a) present the time of generation and integration of a varied sequences of triples over an empty RDF document. When the percentage of insertions in the sequence is 100% the performance of our algorithm increases. This is due to the minimal causality dependency between insertions operations computed during the local generation of triples. The Figure 4(b) illustrate the time of generation and integration of a varied sequences of triples over a RDF document containing 10 000 triples. The performance decreases when the percentage of deletion increases. This degradation of performance is caused by the canonizing of the log [5] (tidy insertion operation before deletion operations). The rate of deletion operations in the log has a direct impact on the performance of the Optic algorithm.

5 Conclusion

In this paper, we have dealt with the problem of the real-time co-authoring of LDW. In this respect, we have suggested two solutions based on OT approach.

In centralized and decentralized solutions we propose in this paper, the shared RDF graph is serialized into a sequence of triples that can be altered by simple operations: insertion and deletion of triples. Mapping RDF graph into sequence of triples is given in order to reuse state-of-the art collaborative editing techniques including some systems in which we participated [9,6]. This mapping is simple. But, if the RDF graph must satisfy some requirements based on semantic aspects (*e.g.*, graph connectiveness), preconditions must be added to operations. For example, we can state that the delete operation $Del(p, t)$ is enabled iff the p exits and the object of t is not a subject of another triple. It is not sure that this delete operation will be still enabled when it is integrated in another site which has added concurrently triple t' whose the subject is the object of t. Two solutions are possible: either writing another IT algorithm based on new constraints, or tolerating the violation of some requirements during some periods with the possibility to stabilize in correct state (by undoing some operations).

The question of adapting these solutions in existing semantic web browsers remains open in this paper. It will be interesting to plug these solutions in a given browser in order to evaluate the cost of mapping a RDF graph into a sequence. Using this implementation, we can also make measurements to experimentally validate the impact of OT approach on real-timeliness and scalability. On the other hand, designing a new IT algorithm for shared RDF graphs based on updates proposed in the recent version of SPARQL/Update [12] is an exciting and challenging problem.

References

1. Aslan, K., Molli, P., Skaf-Molli, H., Weiss, S.: C-set: a commutative replicated data type for semantic stores. In: Fourth International Workshop on REsource, RED (2011)

2. Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, J., Hollenbach, R., Lerer, A., Sheets, D.: Tabulator: Exploring and analyzing linked data on the semantic web. In: SWUI06 Workshop at ISWC 2006, Athens, Georgia, USA (2006)
3. Berners-Lee, T., Hollenbach, J., Kanghao, L., Presbery, E., Pru d'ommeaux, J., Schraefel, M.: Tabulator redux: Writing into the semantic web (2008), http://eprints.ecs.soton.ac.uk
4. Ellis, C.A., Gibbs, S.J.: Concurrency Control in Groupware Systems. In: SIGMOD Conference, vol. 18, pp. 399–407 (1989)
5. Imine, A.: Conception Formelle d'Algorithmes de Réplication Optimiste. Vers l'Edition Collaborative dans les Réseaux Pair-à-Pair. Phd thesis, University of Henri Poincaré, Nancy, France (December 2006)
6. Imine, A., Rusinowitch, M., Oster, G., Molli, P.: Formal design and verification of operational transformation algorithms for copies convergence. Theoretical Computer Science 351(2), 167–183 (2006)
7. Li, R., Li, D., Sun, C.: A time interval based consistency control algorithm for interactive groupware applications. In: IEEE ICPADS 2004, Los Alamitos, CA, USA, pp. 420–429 (2004)
8. Manola, F., Miller, E.: Rdf primer (2004), http://www.w3.org/TR/rdf-primer/
9. Molli, P., Oster, G., Skaf-Molli, H., Imine, A.: Using the transformational approach to build a safe and generic data synchronizer. In: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, pp. 212–220. ACM Press (2003)
10. Ressel, M., Nitsche-Ruhland, D., Gunzenhauser, R.: An Integrating, Transformation-Oriented Approach to Concurrency Control and Undo in Group Editors. In: ACM CSCW 1996, Boston, USA, pp. 288–297 (November 1996)
11. Sacco, O., Collina, M., Schiele, G., Corazza, G.E., Breslin, J.G., Hauswirth, M.: Fine-grained access control for RDF data on mobile devices. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) WISE 2013, Part I. LNCS, vol. 8180, pp. 478–487. Springer, Heidelberg (2013)
12. Seaborne, A., Manjunath, G.: Sparql/update: A language for updating rdf graphs (2008), http://jena.hpl.hp.com/~afs/SPARQL-Update.html
13. Sun, C., Agustina.: Exhaustive search of puzzles in operational transformation. In: CSCW 2014, pp. 519–529 (2014)
14. Sun, C., Ellis, C.: Operational transformation in real-time group editors: issues, algorithms, and achievements. In: ACM CSCW 1998, Seattle, Washington, United States, pp. 59–68 (1998)
15. Sun, C., Jia, X., Zhang, Y., Yang, Y., Chen, D.: Achieving Convergence, Causality-preservation and Intention-preservation in real-time Cooperative Editing Systems. ACM Trans. Comput.-Hum. Interact. 5(1), 63–108 (1998)
16. Vidot, N., Cart, M., Ferrié, J., Suleiman, M.: Copies convergence in a distributed real-time collaborative environment. In: ACM CSCW 2000, Philadelphia, USA (December 2000)
17. Yi, X., Chengzheng, S., Mo, L.: Achieving convergence in operational transformation: Conditions, mechanisms and systems. In: CSCW 2014, pp. 505–518 (2014)
18. Zander, S., Schandl, B.: Context-driven RDF data replication on mobile devices. Semantic Web 3(2) (2012)

Software Engineering: Modeling and Meta Modeling

A High Level Net for Modeling and Analysis Reconfigurable Discrete Event Control Systems

Ahmed Kheldoun[1,4(✉)], Kamel Barkaoui[2], JiaFeng Zhang[3], and Malika Ioualalen[1]

[1] MOVEP, Computer Science Department, USTHB, Algiers, Algeria
ahmedkheldoun@yahoo.fr, mioualalen@usthb.dz
[2] CEDRIC-CNAM, 292 Rue Saint-Martin 75141, Cedex 03, Paris, France
kamel.barkaoui@cnam.fr
[3] School of Electro-Mechanical Engineering, Xidian University, Xi'an, 710071, China
zhangjiafeng628@gmail.com
[4] Sciences and Technology Faculty, Yahia Fares University, Medea, Algeria

Abstract. This paper deals with automatic reconfiguration of discrete event control systems. We propose to enrich the formalism of recursive Petri nets by the concept of $feature$ from which runtime reconfigurations are facilitated. This new formalism is applied in the context of automated production system. Furthermore, the enhanced recursive Petri net is translated into rewriting logic, and by using Maude LTL model-checker one can verify several behavioural properties related to reconfiguration.

Keywords: Reconfigurable control systems · Feature · Recursive Petri nets · Rewriting logic · Maude

1 Introduction

The new generation of discrete event control models is adressing new creteria as flexibility and agility. The need of flexibility and adaptability leads to integrate reconfigurability features in these models, but it makes the system more complex and its development a hard task. Therefore, an approach for the design safe and reconfigurable systems is a crucial need. The Petri net formalism is one of the most used tools to model and analyse discrete event systems [2].

Recently, recursive Petri nets (RPNs) [3] are proposed to specify flexible concurrent systems where functionalities of discrete event systems such as abstraction, dynamicity, preemption, recursion are preponderant. In fact, RPNs have ability to model dynamic creation of threads which behave concurrently.

In this paper, we introduce the concept of $feature$ proposed in [13] to deal with reconfiguration at runtime. More precisely, the reconfiguration is modelled by combining the interruption and the activation/deactivation of transitions which is ensured by : $application\ condition$ and $update\ expression$.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 recalls the syntax and semantic of the formalism RPNs. The formalism which enrich RPN by the concept of $feature$, named reconfigurable RPN and denoted by R^2PN, is presented in Section 4. Section 5 presents a case study of a

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 551–562, 2015.
DOI: 10.1007/978-3-319-19578-0_45

reconfigurable automated production system, and we present in Section 6 its modelling in terms of R^2PN. The verification of the obtained model is done by using the LTL model-checker of Maude [6] [11] and is described in Section 7. Section 8 concludes this paper and depicts further research work.

2 Related Work

Many researchers have tried to deal with formal modeling of control systems with potential reconfigurations. The author of [1] proposed self-modifying nets that can modify their own firing rules at runtime, however, most of the net basic properties such as reachability, boundedness and liveness become indecidable on these nets. In [4], the authors developped a Reconfigurable Petri Nets (RPN) for modeling adaptable multimedia and protocols that can self-modify during execution. They modelled the reconfiguration by introducing the concept of *modifier* places. The authors of [5] presented net rewriting systems (NRS) where a reconfiguration of the net is obtained by a rewriting rules execution. The rewriting rules are similar to production of graph grammars. However, the formalism of NRS is Turing powerful and, thus, automatic verification is no longer possible in that case. Recently, in [7], the authors proposed Reconfigurable timed net condition/event systems (R-TNCES) for modeling reconfigurable discrete event control systems. In this formalism, the system is represented by a set of control componnents and a reconfiguration is modelled by enabling/disabling some control components modules by changing condition/event signals among them.

In this paper, we present a new formalism named Reconfigurable RPN (R^2PN) enriches RPN by the concept of feature selection introduced in [13]. Indeed, in R^2PN, the reconfiguration is modelled by combining the interruption and refinement with the activation/deactivation of transitions which is ensured by : $application\ condition$ and $update\ expression$. Moreover R^2PN captures the behaviour of entire reconfigurable discrete event control system in a concise modular model, opening the way for efficient analysis and verification.

3 Recursive Petri Nets

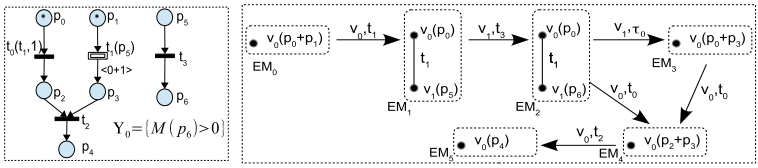
The formalism of RPN [3] consider two types of transitions : elementary and abstract. Moreover a starting marking is associated to each abstract transition and a semi-linear set of final markings is defined.

Definition 1. *(Recursive Petri Nets). A Recursive Petri Net [3] is defined by a tuple* $N = \langle P, T, Pre, Post, \Omega, I, \Upsilon, K \rangle$ *where:*

- P *is a finite set of places.*
- T *is a finite set of transitions where* $T = T_{el} \uplus T_{abs}$ *named respectively, the set of elementary and abstract transitions,*
- I *is a finite set of indices called termination indices,*
- Pre *is a mapping defined as :* $Pre : T \rightarrow P^{\oplus}$*, where* $P^{\oplus}$ *is the set of finite multi-sets over the set* P*,*

- *Post is a mapping defined as :* $Post : T_{el} \cup (T_{abs} \times I) \rightarrow P^{\oplus}$,
- Ω *is a mapping* $T_{abs} \rightarrow P^{\oplus}$ *associating to each abstract transition an ordinary marking,*
- Υ *is a family indexed by* I *of termination sets, where each set represents a set of final markings (i.e. un element of* $P^{\oplus}$*),*
- $K : T_{el} \rightarrow T_{abs} \times I$*, maps a set of interrupted abstract trasitions, and their associated termination indexes, for every elementary transition.*

Example 1. Let's use the net presented in Fig.1(a) to highlight RPN's graphical symbols and associated notations. (i) An elementary transition is represented by a filled rectangle; its name is possibly followed by a set of terms $(t', i) \in T_{abs} \times I$. Each term specifies an abstract transition t', which is under the control of t, associated with a termination index to be used when aborting t' consequently to a firing of t. For instance, t_0 is an elementary transition where its firing preempts threads started by the firing of t_1 and the associate index is 1. (ii) An abstract transition t is represented by a double border rectangle; its name is followed by the starting marking $\Omega(t)$. For instance, t_1 is an abstract transition and $\Omega(t_1) = p_5$ means that any thread, named refinement net, created by firing of t_1 starts with one token in place p_5. (iii) Any termination set can be defined concisely based on place marking. For instance, Υ_0 specifies the final marking of threads such that the place p_6 is marked at least by one token. (iv) The set I of termination indices is deduced from the indices used to subscript the termination sets and from the indices bound to elementary transitions i.e. interruption. In this example, $I = \{0, 1\}$.



(a) A recursive Petri net N (b) Two firing sequences of N

Fig. 1. A recursive Petri net and Two possible firing sequences

Informally, a RPN generates during its execution a dynamical tree of marked threads called an extended marking, which reflects the global state of a such net. This latter denotes the fatherhood relation between the generated threads (describing the inter-threads calls). Each of these threads has its own execution context.

Definition 2. *(Extended Marking). An extended marking [3] of a recursive Petri net is a labelled tree*
$EM = \langle V, M, E, A \rangle$ *where:*

- V *is the (possibly empty) finite set of nodes. When it is non empty,* v_0 *denotes the root of the tree,*

- *M is a mapping $V \rightarrow P^{\oplus}$ associating an ordinary marking for each node,*
- *$E \in V \times V$ is the set of edges,*
- *A is a mapping $E \rightarrow T_{abs}$ associating an abstract transition for each edge.*

Any ordinary marking can be seen as an extended marking composed by a unique node. The empty tree is denoted by $\perp$. Note contrary to ordinary nets, RPNs are often disconnected since each connected component may be activated by the firing of abstract transitions. In a RPN, we have two kinds of markings: extended markings and ordinary markings. An extended marking represents the state of the RPN. An ordinary marking represents an execution context of the thread as in Petri nets.

Definition 3. *(Enabled transition or cut step [3]).*

- *A transition is enabled in a node v of an extended marking $EM \neq \perp$ denoted by $EM \xrightarrow{v,t}$ if $\forall p \in P : M(v)(p) \geq Pre(p,t)$,*
- *A cut step τ_i is enabled in a node v if $M(v) \in \Upsilon_i$.*

The firing of an elementary transition updates the current marking using ordinary firing rule like in Petri nets. The firing of an abstract transition refines it by a new sub-net (i.e. creation of new thread, named its child) which starts its own token game, from a starting marking whose value is attached to the abstract transition. Once a final marking is reached, a cut step closes the corresponding sub net, kills its children and produces tokens, indicated by the $Post$ function, in the appropriate output places of the abstract transition. Formal definitions of firing rules are defined in [3]. Due to lack of space, we explain their principls through our illustrated example of Fig.1(a).

Example 2. Fig.1(b) highlights a firing sequences of RPN represented in Fig.1(a). The graphical representation of any extended marking EM is a tree where an arc $v_i(m_i) \xrightarrow{t_{abs}} v_j(m_j)$ means that v_j is a child of v_i created by firing the abstract transition t_{abs} and m_i (reps. m_j) is the marking of v_i (reps. v_j). Note that the initial extended marking EM_0 is reduced to a single node v_0 whose marking is $p_0 + p_1$. From the intial extended marking EM_0, the abstract transition t_1 is enabled; its firing leads to the extended marking EM_1 which contains a fresh node v_1 marked by the starting marking $\Omega(t_1)$. Then, the firing of the elementary transition t_3 from node v_1 of EM_1 leads to an extended marking EM_2, having the same structure as EM_1 but only the marking of node v_1 is changed. From node v_1 in EM_2, the cut step τ_0 is enabled; its firing leads to an extended marking EM_3 by removing the node v_1 and change the marking on its node predecessor i.e. v_0 by adding $Post(p_3, t_1, 0)$. Also, another way to remove nodes in extended marking is using the concept of preemption associated to the elementary transitions. For instance, from node v_0 in EM_2, the elementary transition t_0 with associated preemption $(t_1, 1)$ is enabled; its firing leads to an extended marking EM_4 by removing the node v_1.

4 Reconfigurable Recursive Petri Nets

Reconfigurable Recursive Petri nets (R^2PNs) enriches RPN by the concept of feature selection introduced in [13]. In fact, R^2PNs extend RPN by associating transitions and cut steps with application conditions and update expressions. An application condition

is a logical formula over a set of features, describing the feature combinations to which the transition applies. It consitutes a necessary (although not sufficient) condition for the transition to fire. In fact, if the application condition is false, means that the transition is desactivated. An update expression, describes the feature selection evolves after firing a transition.

A feature is defined as a prominent or distinctive user-visible aspect, quality or characteristic of a system. A feature is defined in [13] as follows :

Definition 4. *(Feature [13]). A feature is an end-user visible characteristic of a system.*

The concept of feature has been introduced by the software design community to specify and distinguish products in product lines [9][13]. Now, let's define the set of application conditions over a set of features.

Definition 5. *(Application condition). An application condition φ [9] is a logical (boolean) contraint over a set of features F, defined by the following grammar: $\varphi ::= true \mid a \mid \varphi \wedge \varphi \mid \neg\varphi$, where $a \in F$. The remaining logical connectives can be encoded as usual. We write Φ_F to denote the set of all application conditions over F.*

Definition 6. *(Satisfaction of application conditions [9]). Given an application condition φ and a sub set of features FS, called a feature selection, we say that FS satisfies φ, written as $FS \models \varphi$, iff: (1) $FS \models true$ always; (2) $FS \models a$ iff $a \in FS$; (3) $FS \models \neg\varphi$ iff $FS \not\models \varphi$; (4) $FS \models \varphi_1 \wedge \varphi_2$ iff $FS \models \varphi_1$ and $FS \models \varphi_2$*

Definition 7. *(Update). An update[9] is defined by the following grammar: $u ::= noop \mid a\ on \mid a\ off \mid u; u$, where $a \in F$ and F is a set of features. We write U_F to denote the set of all updates over F. Given a feature selection $FS \subseteq F$, an update expression modifies FS according to the following rules:* r1: $FS \xrightarrow{noop} FS$; r2: $FS \xrightarrow{a\ on} FS \cup \{a\}$; r3: $FS \xrightarrow{a\ off} FS \setminus \{a\}$; r4: $\frac{FS \xrightarrow{u_0} FS' \quad FS' \xrightarrow{u_1} FS''}{FS \xrightarrow{u_0;u_1} FS''}$.

We are now in position to introduce R^2PNs.

Definition 8. *(Reconfigurable Recursive Petri nets). A R^2PNs is a tuple $EN = \langle N, F, f, u\rangle$, where :*

- *$N = \langle P, T, Pre, Post, \Omega, I, \Upsilon, K\rangle$ is RPN,*
- *F is a set of features,*
- *$f : T \cup \{\tau_i\} \to \Phi_F$ is a function associating to each transition and cut step with an application condition from Φ_F where $i \in I$,*
- *$u : T \cup \{\tau_i\} \to U_F$ is a function associating to each transition and cut step with an update from U_F where $i \in I$.*
  We write u_t resp. u_{τ_i} to denote the update expression $u(t)$ resp. $u(\tau_i)$ associated to a transition t resp. a cut step τ_i.

we write $FS \models f(t)$ if the feature selection FS satisfies the application condition associated with transition t. In the following, graphically, each transition of R^2PN is annotated by an application condition and an update expression in the following way:

$$\frac{application\ condition}{update\ expression}$$

Definition 9. *(A state of Reconfigurable RPN). A state of a Reconfigurable RPN* $EN = \langle N, F, f, u\rangle$ *is a tuple* $S = (EM, FS)$ *where* $EM = \langle V, M, E, A\rangle$ *is an extended marking and* $FS \subseteq F$ *is a feature selection.*

Definition 10. *(Enabled transition or cut step). Let* $S = (\langle V, M, E, A\rangle, FS)$ *be a state of* R^2PN $EN = \langle N, F, f, u\rangle$ *where* $N = \langle P, T, Pre, Post, \Omega, I, \Upsilon, K\rangle$. *Let a node* $v \in V$.

- *A transition* t *is enabled in a node* v, *if* $\forall p \in P : M(v)(p) \geq Pre(p,t)$ *and* $FS \models f(t)$,
- *A cut step* τ_i *is enabled in a node* v, *if* $M(v) \in \Upsilon_i$ *and* $FS \models f(\tau_i)$.

Definition 11. *(Firing rules of Reconfigurable RPN). Let* $S = (EM, FS)$ *be a state of* R^2PN $EN = \langle N, F, f, u\rangle$ *where* $N = \langle P, T, Pre, Post, \Omega, I, \Upsilon, K\rangle$. *Let a node* $v \in V$.

- *The firing of an elementary transition* t *from a node* v *leads to a state* $S' = (EM', FS')$ *where* $EM \xrightarrow{v,t} EM'$ *as* Definition12. *in [3]. and* $FS \xrightarrow{u(t)} FS'$,
- *The firing of an abstract transition* t *from a node* v *leads to a state* $S' = (EM', FS')$ *where* $EM \xrightarrow{v,t} EM'$ *as* Definition13. *in [3]. and* $FS \xrightarrow{u(t)} FS'$,
- *The firing of a cut step* τ_i *from a node* v *leads to a state* $S' = (EM', FS')$ *where* $EM \xrightarrow{v,\tau_i} EM'$ *as* Definition14. *in [3]. and* $FS \xrightarrow{u(\tau_i)} FS'$.

Therefore, the analysis of R^2PN is based on constructing its extended reachability graph, which is used for checking properties such as reachability, deadlock and liveness.

5 Case Study : Automated Production Systems

In this research work, we use a reconfigurable production devices called, FESTO[7] as a running example. We assume that the device may perform some particular reconfiguration scenarios according to well-defined conditions. FESTO is composed of three units: distribution, test and processing units. The distribution unit is composed of a pneumatic feeder and a converter to forward cylindrical work pieces from a stack to the testing unit which is composed of the detector, the tester and the elevator. The testing unit checks of work pieces for height, material type and color. Work pieces that successfully pass this check are forwarded to the rotating disk of the processing unit, where the drilling of the work piece is performed. We assume in this work two drilling machines $Dr1$ and $Dr2$ to drill pieces. The result of the drilling operation is next checked by a checking machine and the work piece is forwarded to another mechanical unit. Three production modes (called local configurations) can be performed by $FESTO$.

- $Light1$: For this production mode, only the drilling machine $Dr1$ is used;
- $Light2$: To drill work pieces for this production mode, only the drilling machine $Dr2$ is used;
- $High$: For this production mode, where $Dr1$ and $Dr2$ are used at the same time in order to accelerate the production.

$Ligth1$ is the default production mode of $FESTO$ and the system completely stops in the worst case if the two drilling machines are broken. We assume that $FESTO$ may perform four reconfiguration scenarios as shown in Fig.2.

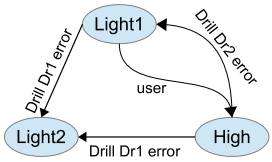


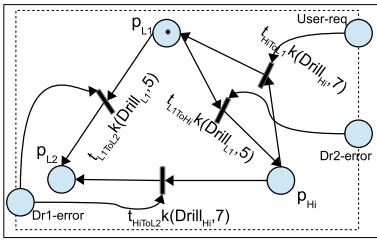
Fig. 2. Reconfiguration scenarios of $FESTO$

6 Modeling $FESTO$ Using Reconfigurable RPN

The automated production system $FESTO$ is modelled as follows: $EN_{FESTO} = \langle EN_{Beh}, EN_{Adapt} \rangle$ where EN_{Beh} represents the behaviour module of $FESTO$ and EN_{Adapt} is the adaptor which represents possible reconfiguration scenarios may be applied by the reconfigurable control system $FESTO$.

The adaptor EN_{Adapt} of $FESTO$ is shown in Fig.3. It is represented by ERPN where each place specifies one beahviour. As shown in Fig.3, we have three places p_{L1}, p_{L2} and p_{Hi} which specify the three production modes $Light1$, $Light2$ and $High$. Each one of these places may contain at most one token and the marking of such place means that its associated production mode is currently applied by the production system $FESTO$. For instance, the place p_{L1} is marked, which means the current production mode applied by $FESTO$ is $Light1$ i.e.the initial production mode. The set of elemenatry transitions represent the set of reconfiguration scenarios of $FESTO$. For instance, the elementary transition t_{L1ToL2} models the reconfiguration scenario that allows the production system $FESTO$ to transform from the first production mode $Light1$ to the second production mode $Light2$ when drilling machine $Dr1$ is broken. In fact, the firing of this transition will interrupt the abstract transition $Drill_{L1}$, which models the first production mode $Light1$, and update the current feature selection FS by applying its associated update expression $Dr1\ off; Dr2\ on$ as shown in Fig.3.

The behaviour EN_{Beh} of $FESTO$ which is a union of multiple R^2PNs is formalised as follows: $EN_{Beh} = \bigcup_{i \in \overline{1..3}} EN_{Beh_i}$, with $EN_{Beh_i} = \langle P_i, T_i, Pre_i, Post_i, \Omega_i, I_i, \Upsilon_i, K_i, F_i, f_i, u_i \rangle$ is a R^2PN models one possible behaviour of reconfigurable control system of $FESTO$. Fig.4 models the behaviour of $FESTO$ using ERPN. All the transitions shown in Fig.4, where their *application condition* and *update expression* are omitted, are annotated by the term : $\frac{true}{noop}$. This means that this set of transitions are common to all behaviours of $FESTO$. The set of features F contains the set of drilling machines which may be used to select the proper behaviour of $FESTO$ i.e. $F = \{Dr1, Dr2\}$. As noted in Fig.4, the abstract transitions $Distribute$, $Test$ and $Process$ models the distribution, tester and processing unit. The firing of one of these transitions will create a thread representing the behaviour of associating unit.



Transition	Application condition	Update condition
t_{L1ToL2}	$Dr_1 \wedge \neg Dr_2$	$Dr_1 off\ ; Dr_2 on$
t_{L1ToHi}	$Dr_1 \wedge \neg Dr_2$	$Dr_2 on$
t_{HiToL1}	$Dr_1 \wedge Dr_2$	$Dr_2 off$
t_{HiToL2}	$Dr_1 \wedge Dr_2$	$Dr_1 off$

Fig. 3. R^2PN represents FESTO's adaptor

For instance, the firing of the abstract transition $Process$ creates a thread, models the behaviour of processing unit, which starts by one token i.e. workpiece in place p_{12}. The workpiece is then forwarded to the drilling machines by firing the elementary transition $Rotate$. After, three abstract transitions $Drill_{L1}$, $Drill_{L2}$ and $Drill_{Hi}$ may be enabled; they model the drilling's step according to the three production modes $Light1$, $Light2$ and $High$. But each one of these abstract transitions is associated an application condition which restricts its activation (firing) to the set of bound features F. As described above, the default production mode of $FESTO$ is $Light1$, where only the drilling machine $Dr1$ is used, so the initial feature selection $FS_0 = \{Dr1\}$. In this case, only the abstract transition $Drill_{L1}$ is enabled. The firing of this abstract transition will create a thread, models the drilling's step, which starts by one token in place p_{17}. Note that the created thread can use only the drilling machine $Dr1$ represented by the elementary transition $Dr1$-$L1$. Moreover, this thread presents two types of termination :

- $Properly\ termination$: it means that the workpiece is well drilled and the place p_{18} is marked. So, a final marking belongs to termination's set Υ_4 is reached, then the cut step τ_4 may be enabled. The firing of τ_4 terminates the current thread and puts a workpiece in the place p_{14} in order to perform the remains operations such as $Checker$ and $Evacuate$.
- $Termination\ by\ interruption$: this termination occured when the production system $FESTO$ applies a reconfiguration as described above for adaptor module. For instance,from Fig.3, firing the elementary transition t_{L1ToL2} will interrupt the thread created by the absract transition $Drill_{L1}$ with termination index 5 and update the feature selection FS_0. The new obtained feature selection $FS_1 = \{Dr2\}$. In fact, the workpiece is put it in the place p_{13} and only the abstract transition $Drill_{L2}$ may be enabled, which specify the drilling's step according to the second production mode $Light2$.

7 Verification of Reconfigurable Control Systems

In this section, we outline the conversion from R^2PNs to a Maude specification [6] and the use of its Linear Temporal Logic (LTL) model checker [11].

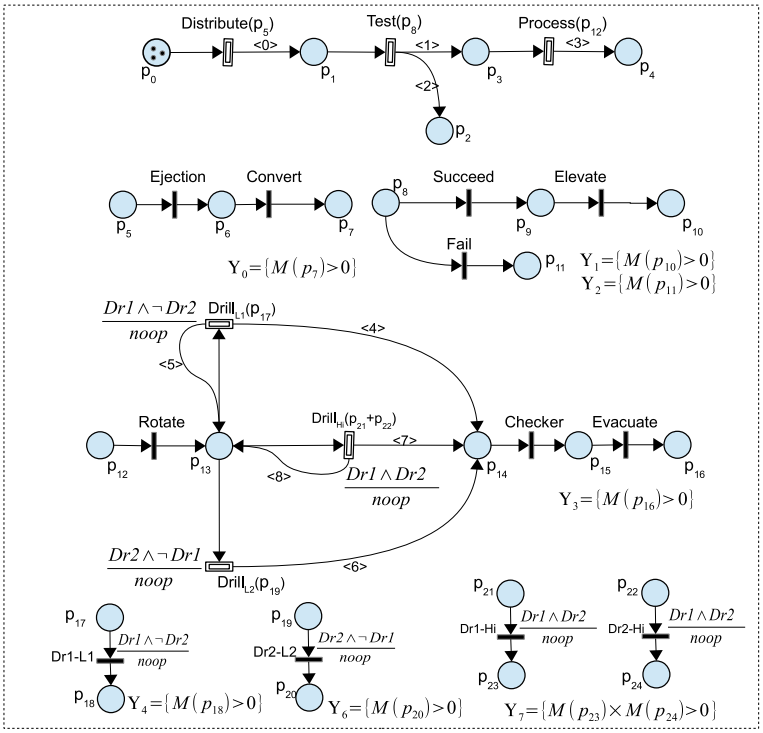


Fig. 4. R^2PN represents FESTO's behaviour

7.1 Maude and Its Model-Checker

Maude is a high-performance reflective language and system supporting rewriting logic specification [12]. It has been developed at SRI (URL: http://maude.cs.uiuc.edu/) International for over two decades. A system, under Maude, is represented using membership equational logic describing its set of states and a set of rewrite rules representing its state transitions. Maude is stricly typed, where the types are called $sorts$ and can be built hierarchically using $subsorts$. Maude's basic programming statements are equations and rules, and have in both cases a simple rewriting semantics in which instances of the left-hand side pattern are replaced by corresponding instances of the right-hand side. One aim using Maude is its LTL model-checker which can be used to verify properties as reachability, deadlock or liveness for a specified model. Model checking can be used to prove properties, specified in LTL when the set of states reachable from an initial state in a system module is finite. In [11], the author presents more details about syntax and semantic of LTL.

7.2 Conversion of R^2PN to a Maude specification

Like in [8], the state of a R^2PN is described by a term $State(EM, fs)$ of sort $STATE$ where:

- EM is an extended marking represented, in a recursive way, as a dynamical tree by the term $[M_{Th}, tabs, ThreadChilds]$ of sort $Thread$ where M, of sort $Marking$, represents the internal marking of Th. The term $tabs$ represents the name of the abstract transition whose firing (in its thread father) gave birth to the thread Th. Note that the root thread is not generated by any abstract transition, so the abstract transition which gave birth to it, is represented by the constant $nullTrans$. The term $ThreadChilds$ represents a finite multiset of threads generated by the firing of abstract transitions in the thread Th. We denote by the constant $nullThread$, the empty thread.
- fs is a feature selection, of sort FS, represented by a list of terms of sort $Term$. We denote by the constant $empty$, the empty list of feature selection.

We have also impleneted two functions in the module $FeatureSel$ needed by our formalism R^2PN. The first function is $SATAC(ac : AC, fs : FS) : Bool$ which checks the truth value of an application condition ac, of sort AC, for a given feature selection fs. The second function is $UPDATE(u : UE, fs : FS) : FS$ which returns the new feature selection after applying the update expression u, of sort UE, for a given feature selection fs.

Moreover, each transition firing and cut step execution is formally specified in Maude by a labelled rewrite rule as follows :

- Rule associated to an elemetary transition t with $K(t) = \phi$, application condition $ac(t)$ and update expression $ue(t)$

```
crl[t]: State(<p; N+Pre(p,t)> (*) <p'; M> , fs) =>State(<p; N
    )> (*) <p'; M + Post(p',t)> , UPDATE(ue(t), fs) if SATAC(
    ac(t), fs).
```

- Rule associated to an elemetary transition t with $K(t) = \{(t_{absi}, k), (t_{absj}, m), ..\}$, application condition $ac(t)$ and update expression $ue(t)$

```
crl[t]: State([<p;N+Pre(p,t)>(*)<p';M>(*)<p'_i;A>(*)<p'_j;B>,
    absTrans,Thread],fs) => State([<p; N>(*)<p'; M+Post(p',t)
    >(*)<p'_i; A+Post(p'_i, t_absi,k)>(*)<p'_j; B+Post(p'_j,t_absj,m)>,
    absTrans, DeleteThread(t_absi,t_absj,...,Thread)], UPDATE(ue(t
    ), fs) if SATAC(ac(t), fs).
```

- Rule associated to an abstract transition t with starting marking $\Omega(t)$, application condition $ac(t)$ and update expression $ue(t)$

```
crl[t]: State([<p;N+Pre(p,t)>, absTrans, Thread] , fs) =>
    State([<p; N)>, absTrans, Thread[<p'; Ω(t)>, t, nullThread
    ]] , UPDATE(ue(t), fs) if SATAC(ac(t), fs).
```

- Rule associated to a cut step τ_i with application condition $ac(\tau_i)$ and update expression $ue(\tau_i)$

```
crl[τ_i]: State([<p;N>, absTrans, Thread[<p';N'> , tabs,
    Thread1]] , fs) => State([<p; N+Post(p, tabs, i)>,
    absTrans, Thread, UPDATE(ue(τ_i), fs) if (Υ_i and SATAC(ac(τ_i
    ), fs)).
```

7.3 Implementation Using the Maude Tool

Since we give a Maude specification for the formalism R^2PN, we can benefit from the use of the LTL model-checker of the Maude system for verification purpose where the generated state space must be finite. For instance, one can check the liveness property over EN_{FESTO} for its initial behaviour $Ligth1$. We suppose that the system starts by100 tokens i.e. workpieces, this is specified in Maude by : $eqinitialState = State(< p0; 100 > (*) < pL1; 1 >, Dr1)$. A liveness condition is : *each workpiece must reach (from all reachable markings) the final state where the place* p_4 *is marked*.This can be phrased as "*For all paths and from all states,* $State(< p4; 100 > (*) < pL1; 1 >, Dr1)$ *can finally be reached*". In Maude, this is stated by $[] <> State([< p4; 100 > (*) < pL1; 1 >, nullTrans, nullThread], Dr1)$., and proven to be $valid$ by its model checker in Fig.5(a). We suppose in this case that there is no $fail$ during the workpieces's test process.

Let take another example and we focus on the case, when an error occurs, whether the control module can respond and select a proper behaviour. We define the following LTL formula : $\alpha : [](Behaviour(Light1)/\ Drill - Down(Dr1) => <> Behaviour(Light2)))$, where, the predicate $Behaviour$ allows to know the current behaviour applied by the production system $FESTO$. The predicate $Drill\text{-}Down$ indicates which among drilling machines $Dr1$ ord $Dr2$ is break-down.

This LTL formula means that, always, if the current production mode of $FESTO$ is $Light1$, drill machine $Dr1$ is broken, the production system $FESTO$ will eventually select the production mode $Light2$. This LTL formula is proved to be $valid$ in Fig.5(b).

Now, let's define a LTL proprty β by replacing in the formula α the production mode $Light2$ by $High$. In Fig.5(c), this formula is proved to be $not\ valid$ and the model-checker returns the expected $counterexample$.

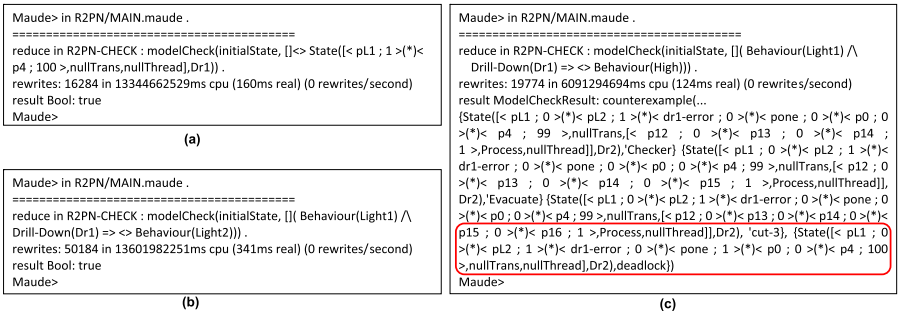


Fig. 5. **(a)** Model checking of the liveness condition for first production mode of $FESTO$, **(b)** Model checking of the LTL property α and **(c)** $Counterexample$ generated by model checking of the LTL property β

8 Conclusion and Future Work

This research work copes with the reconfiguration issue of discrete control systems. We have proposed Renconfigurable RPN (R^2PN) which enriches RPN by the concept of $feature$ to deal with reconfigurations at runtime. R^2PN allows instance of threads in RPN to be renconfigurable. We have shown the efficiency of R^2PN through a case study represented by a reconfigurable production system. A verification method for R^2PN has also been presented by using the LTL model-cheker of Maude.

In the future, we will plan to extend our formalism in order to model time constraints which are of great importance in real-time systems. Therefore, one can verify some properties with respect to time constraints using Real-Time Maude model-checker [10].

References

1. Valk, R.: Self-modifying nets, a natural extension of petri nets. In: Proceedings of the Fifth Colloquium on Automata, Languages and Programming, pp. 464–476 (1978)
2. Murata, T.: Petri nets: Properties, analysis and applications. Proceedings of the IEEE 77(4), 541–580 (1989)
3. Haddad, S., Poitrenaud, D.: Recursive Petri nets – Theory and application to discretee event systems. Acta Informatica 44(7-8), 463–508 (2007)
4. Guan, S.-U., Lim, S.-S.: Modeling adaptable multimedia and self-modifying protocol execution. Future Gener. Comput. Syst. 20(1), 123–143 (2004)
5. Badouel, M.L.E., Oliver, J.: Modeling concurrent systems: Reconfigurable nets. In: Proc. Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA 2003), pp. 1568–1574. CSREA Press (2003)
6. Clavel, M., Duran, F., Eker, S., Lincoln, P., Marti-Oliet, N., Meseguer, J., Quesada, J.: Maude: specification and programming in rewriting logic. Theoretical Computer Science 285(2), 187–243 (2002), rewriting Logic and its Applications
7. Zhang, J., Khalgui, M., Li, Z., Mosbahi, O., Al-Ahmari, A.: R-tnces: A novel formalism for reconfigurable discretee event control systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems 43(4), 757–772 (2013)
8. Barkaoui, K., Hicheur, A.: Towards analysis of flexible and collaborative workflow using recursive eCATNets. In: ter Hofstede, A.H.M., Benatallah, B., Paik, H.-Y. (eds.) BPM Workshops 2007. LNCS, vol. 4928, pp. 232–244. Springer, Heidelberg (2008)
9. Muschevici, R.: Modelling Diversity in Software Product Lines. PhD thesis, KU Leuven university, Belgium (December 2013)
10. Ölveczky, P.C., Meseguer, J.: Real-Time Maude: A tool for simulating and analyzing real-time and hybrid systems. In: 3rd International Workshop on Rewriting Logic and its Applications (WRLA 2000). Electronic Notes in Theoretical Computer Science, vol. 36 (2000)
11. Eker, S., Meseguer, J., Sridharanarayanan, A.: The maude LTL model checker. Electronic Notes in Theoretical Computer Science 71, 162–187 (2004)
12. Meseguer, J.: Conditioned rewriting logic as a united model of concurrency. Theor. Comput. Sci. 96, 73–155 (1992)
13. Kang, K., Cohen, S., Hess, J., Novak, W., Peterson, A.: Feature-oriented domain analysis (foda) feasibility study. Software Engineering Institute, Carnegie Mellon University, Tech. Rep. CMU/SEI-90-TR-021 (1990)

Hybrid Approach for Metamodel and Model Co-evolution

Fouzia Anguel[1,3(✉)], Abdelkrim Amirat[2], and Nora Bounour[3]

[1] Chadli Bendjedid University, El Tarf, Algeria.
fanguel@yahoo.fr
[2] LiM Laboratory, Mohammed Chérif Messaadia University, Souk-Ahras, Algeria
abdelkrim.amirat@yahoo.com
[1,3] LISCO Laboratory. Badji Mokhtar University.Annaba, Algeria
nora_bounour@yahoo.fr

Abstract. Evolution is an inevitable aspect which affects metamodels. When metamodels evolve, model conformity may be broken. Model co-evolution is critical in model driven engineering to automatically adapt models to the newer versions of their metamodels. In this paper we discuss what can be done to transfer models between versions of a metamodel. For this purpose we introduce hybrid approach for model and metamodel co-evolution, that first uses matching between two metamodels to discover changes and then applied evolution operators to migrate models. In this proposal, migration of models is done automatically; except, for non resolvable changes, where assistance is proposed to the users in order to co-evolve their models to regain conformity.

Keywords: Metamodel evolution · Model migration · Co-evolution · Matching · Evolution operator

1 Introduction

In Model-Driven Engineering (MDE)[1], metamodels and domain-specific languages are key artifacts as they are used to define syntax and semantics of domain models [1]. Since in MDE metamodels are not created once and never changed again, but are in continuous evolution, different versions of the same metamodel are created and must be managed [2]. The evolution of metamodels is a considerable challenge of modern software development as changes may require the migration of their instances. Works in this direction exist already. Several manual and semi-automatic approaches for realizing model migration have been proposed. Each approach aims to reduce the effort required to perform this process. Unfortunately, in several cases it is not possible to automatically modify the models to make them conform to the updated metamodels. This is so because certain changes over metamodels require introducing additional information into the conformant model. In the literature, three general approaches to the migration of models exist: manual, state-based, operator- based [3]. Manual approaches are tedious and error prone. State-based approaches also called difference-based approaches allow synthesizing a model migration based on the

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 563–573, 2015.
DOI: 10.1007/978-3-319-19578-0_46

difference between two metamodel versions. In contrast, operator-based approaches allow to incrementally transforming the metamodel by means of coupled operations which also encapsulate the corresponding model migration. They allow capturing the intended model migration already when adapting the metamodel. A major drawback of the later approach has been overly tight coupling between the tool performing the migration, and the recorder tracking the changes made to the models.

Usually, existing approaches try to find how to best accomplish model co-evolution. Essentially, we can define two main requirements: the correctness of migration and minimizing the effort of migration by automating as far as possible the process.

In this paper, we propose an alternative solution to model migration which combines state-based and operator based principles to co-evolve models and metamodels. Our vision to resolve this problem is to generate evolution strategies with their corresponding model migration strategies. We focus on including users decisions during metamodel and model co-evolution process to ensure semantic correctness of evolved models.

The rest of the paper is structured as follows. Section 2 gives an overview of basic concepts and describes the metamodel and model co-evolution problem Section 3 presents our proposed approach for solving the model co-evolution problem. In section 4, we present some proposed approaches in the past and situates our solution. Section 5 presents some guidelines to implement proposed framework. Finally, section 6 concludes and gives some future works.

2 Background

2.1 Models and Metamodels

In this section we present the central MDE definitions used in this paper. The basic assumption in MDE is to consider models as first-class entities. An MDE system basically consists of metamodels, models, and transformations. A model represents a view of a system and is defined in the language of its metamodel [1]. In other words, a model contains elements conforming to concepts and relationships expressed in its metamodel [4]. A metamodel can be given to define correct models. In the same way a model is described by a metamodel, a metamodel in turn has to be specified in a rigorous manner; this is done by means of meta-metamodels [5]. This may be seen as a minimal definition in support of the basic MDE principle “Everything is considered as a model” [1]. The two core relations associated to this principle are called representation “Represented by” and conformance “Conform To”. A model conforms to a metamodel, when the metamodel specifies every concept used in the model definition, and the models uses the metamodel concepts according to the rules specified by the metamodel [1].

In this respect, the object management group (OMG) [6] has introduced the four level architecture which organizes artifacts in a hierarchy of model layers (M0, M1, M2, and M3). Models at every level conform to a model belonging to the upper level. M0 is not part of the modeling world as depicted in Fig.1, so the four level architecture should more precisely be named (3+1) architecture [1]. One of the best

known metamodels in the MDE is the UML (Unified Modeling Language) metamodel; MOF (Meta-Object Facility) is the metametamodel of OMG that supports rigorous definition of modeling languages as UML [6].

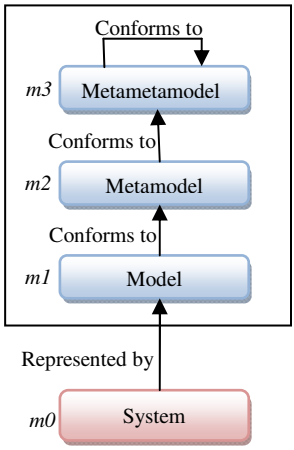


Fig. 1. The 3+1 MDA organisation [2]

2.2 Metamodel Evolution and Model Co-evolution

Metamodels may evolve in different ways, due to several reasons [2]: during design, alternative metamodel versions are developed and well-known solutions are customized for new applications. During implementation, metamodels are adapted to a concrete metamodel formalism supported by a tool. During maintenance, errors in a metamodel are corrected. Furthermore, parts of the metamodel are redesigned due to a better understanding or to facilitate reuse. The addition of new features and/or the resolution of bugs may change metamodels, thus causing possible problems of inconsistency to existing models which conform to the old version of the metamodel and may become not conform to the new version. Therefore to maintain consistency, metamodel evolution requires model adaptation, i.e., model migration; so these two steps are referred as model and metamodel co-evolution [7]. Metamodel and model co-evolution is a term that denotes a coupled evolution of metamodels and models [7], which consists to adapt (co-evolve) the models conforming to the initial version of the metamodel, such that they conform to the target (evolved) version, preserving the intended meaning of the initial model if possible [7], as illustrated in Fig.2. Furthermore, model adaptations should be done by means of model transformations [8]. A model transformation takes as input a model conforming to a given metamodel and produces as output another model conforming to the evolved version of the given metamodel [4].

A number of works proposed the classification of metamodel changes according to their corrupting effects. Metamodel changes are grouped on three categories [9]:

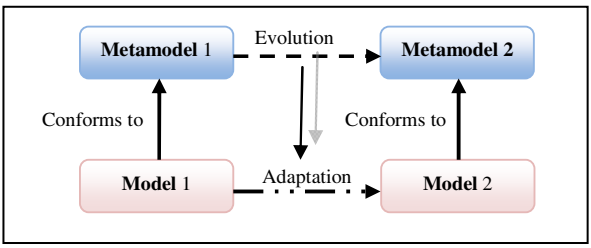


Fig. 2. Model Co-evolution [7]

- Not breaking changes, changes occurring in the metamodel don't break the models conformance to the metamodel.
- Breaking and resolvable changes, changes occurring in the metamodel do break the models, which can be automatically resolved.
- Breaking and non-resolvable changes, changes do break the models and cannot be automatically resolved and user intervention is required.

However, a uniform formalization of metamodel evolution is still lacking. The relation between metamodel and model changes should be formalized in order to allow reasoning about the correctness of migration definitions.

2.3 Logic Programming

Logic programming is a programming paradigm based on formal logic [10]. A program written in a logic programming language is a set of sentences in logical form, expressing facts and rules about some problem domain. Major logic programming language families include Prolog, Answer Set Programming (ASP) and Datalog. In all of these languages, rules are written in the form of clauses (H :- B1, …, Bn). These clauses are called definite clauses or Horn clauses and are read declaratively as logical implications (H if B1 and … and Bn). Logic programming is used in artificial Intelligence knowledge representation and reasoning.

We have find this formalism very powerful to represent relationships between changes and consequently, from an initial set of changes inferring all possible evolution strategies. Currently, to our best knowledge, there is no approach that uses an intelligent reasoning for defining model migrations. Therefore, we have integrated logic programming in our proposal to resolve model co-evolution problem.

3 Proposed Approach

In this section we describe our proposal to ensure the co-evolution of model with their metamodels. The overall evolution and co-evolution process is presented in Fig.3.

Our approach is hybrid because it exports techniques from state-based and operator based approaches and uses also a reasoning mechanism from artificial intelligence. It contains four phases: changes detection, generation and validation of evolution strategies, determination of migration strategies and migration of models.

In the first step; differences between two metamodel versions need to be determined by using matching technique. In the second step we use an inference engine to generate different evolution strategies by assembling atomic changes in possible compound ones; in the third step we explore a library of operators to obtain different migration procedures, which will be assembled to constitute migration strategies. In the last step users employ a selected evolution strategy and consequently, the migration strategy will be applied over a specific model conforming to the old version in order to obtain a new model conforming to the newer metamodel version.

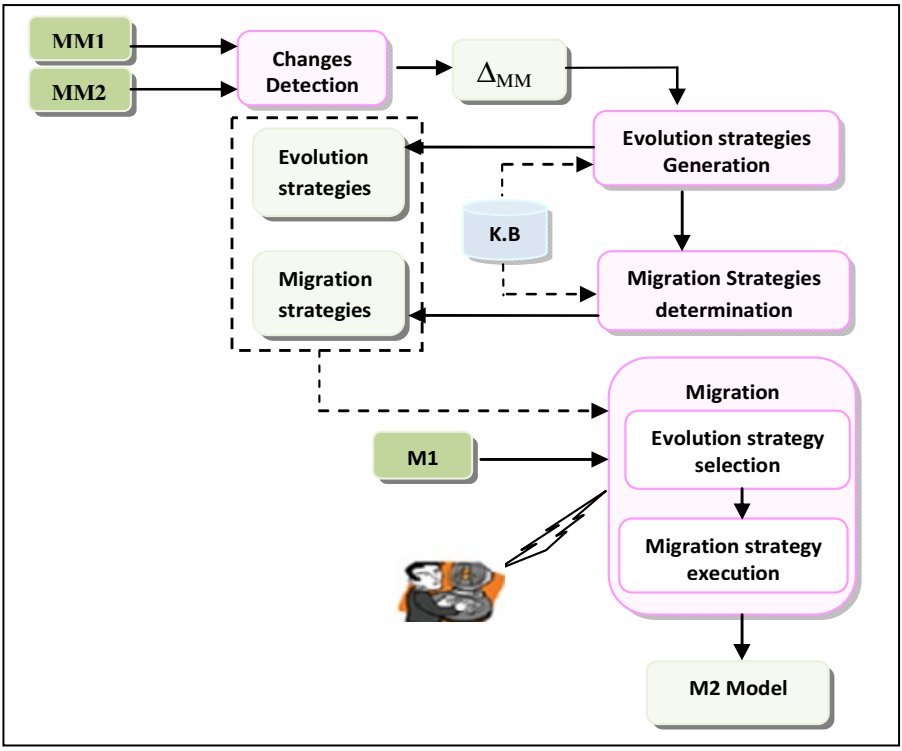


Fig. 3. An overview of metamodel and model co-evolution process

3.1 Detection of Changes

The detection of differences between models is essential to model development and management practices. Thus evolution from one metamodel version to the next can be described by a sequence of changes. Understanding how metamodels evolve or discovering changes that have been performed on a metamodel is a key requirement before undertaking any migration operation on models to co-evolve them. In fact, we distinguish two ways for discovering changes: matching approaches and recording approaches [11]. In Our approach, for detecting the set of changes performed to the older version of the metamodel in order to produce the new one, we use generic

algorithm. Whereas current generic approaches only support detecting atomic changes, some language-specific approaches also allow detecting composite changes; but only for one specific modeling language. Primitive differences between metamodels versions are classified in three basic categories: additions, deletions, and updates of metamodel elements. These differences represent elementary changes (i.e. atomic).

In fact, composite or compound changes have been already considered in previous works like [12,13]. But, we envision tackle the problem differently. We call evolution strategy a possible sequence of changes; here changes are either elementary or composite. Thus, a set of composite changes is inferred from the detected set of atomic changes, by using rules that define composite changes in terms of atomic changes. This mechanism is detailed in the following section.

3.2 Generation and Validation of Evolution Strategies

Detected differences are represented as elementary changes specifying fine-grained changes that can be performed in the course of metamodel evolution. There are a number of primitive metamodel changes like create element, rename element, delete element, and so on. One or more of such primitive changes compose a specific metamodel adaptation However, this granularity of metamodel evolution changes is not always appropriate.

Often, intent of the changes may be expressed on a higher level. Thus, a set of atomic changes can have together the intent of a composite change. For example, generation of a common superclass sc of two classes c1 and c2 can be done through successive applications of a list of elementary changes, such as 'Add_class sc', 'Add_reference from c1 to sc', and 'Add_reference' from c2 to sc. One way to resolve the problem of identifying composite changes is to use operation recording. But, this solution has some drawbacks.

In our proposal, we use logical predicate with the language Prolog. Horn clauses are used to represent knowledge. Therefore, we formally characterize changes. Detected atomic changes are represented as positive clauses (i.e. facts) and composite changes are specified by rules such as Left hand side contains the composite change and the Right hand side contains a set of associated atomic changes. Thus, the applicability of a compound change can be restricted by conditions in the form of rules. According to this principle, we have formalized a knowledge base. The definition of changes is inspired from the literature [7, 14]. The knowledge base is used by the inference engine to generate possible evolution strategies. Finally, evolution strategies must be validated. This step consists of applying each evolution scenario defined by the strategy on the old input version of the metamodel. If it results the newer input version then the tested strategy is valid and it is retained else the strategy in test is rejected. The final output is a set of valid evolution strategies.

C : set of classes
A : set of attributes
Auxiliary predicates

subclasse(s : C, c:C) : s is subclass of c
added_class(c:C) : c is added to the metamodel
added_attribute(a:A,c:C) : a is added to the class c of the metamodel.
deleted_attribute(a:A,c:C) : a is deleted from the class c of the metamodel.
Is_attribute_of(a:A,c:C) : a is an attribute of the class c.
added_supertype ((s : C, c:C) . specialization/generalization reference is added between s and c.
deleted _supertype ((s : C, c:C) : specialization/generalization reference between s and c is removed.
added-reference (r :R, s: C, d:C) : r is an added reference having as source s class and d as target class.
Extract-superclass(sc ,c1,c2 :C) :- added_class(sc: C),
added_supertype ((sc : C, c1:C),
added_supertype ((sc : C, c2:C).

Complex changes are specified through rules. As an instance, we consider extract super class operation where a class is generalized in a hierarchy by adding a new general class and two references to their subclasses.

3.3 Determination of Migration Strategies

In this step we import techniques of operator based-approaches. We use in this phase a library of operators. Thus, we specify a change as an evolution operation. An operation evolution can be either simple or composite and every operation is defined through a set of parameters. We associate to it information about how to migrate corresponding models in response to a metamodel evolution forming a migration procedure. Migration procedure is encoded as a model transformation that transforms a model such that the new model conforms the metamodel undergoing the change. Furthermore, we explicitly specify in migration procedures some assistance specifications for each change requiring additional information from user to solve it. This makes our library different of that used in previous works [13]. The library does not contain evolution steps but only the migration procedure referenced with evolution operation. In our proposal we take from the library migration procedures corresponding to changes in the evolution strategy; after their instantiation, we assemble them to constitute the complete migration strategy which will be associated to the evolution strategy. The final result in this step is a set of couples (evolution strategy, migration strategy) specified to co-evolve input models conforming to the specified metamodel.

3.4 Migration

This phase takes as input an instance model conforming to the initial metamodel. This model is also called user model. To transform the model to newer version of the metamodel, firstly one of available evolution strategies previously inferred is considered. According to the taken evolution strategy associated migration strategy

will be automatically generated and then applied to the input model. For breaking and irresolvable changes, the system assists establish adequate migration procedure by presenting alternative solutions. Additionally, users can provide additional information to complete the change on the model if necessary. For instance, if the new attribute must be initialized, the user must also be requested for the initial value. If the user is satisfied by the resulted model the process is achieved, otherwise he can try again by selecting other proposed evolution strategy and the process continues so that, until user satisfaction or no choice is available.

3.5 Implementation

In this section, we give details and technical choices made to implement a prototype of the proposed framework. As meta-metamodel, we use Ecore from the Eclipse Modeling Framework (EMF) [16]. However, our approach is not restricted to Ecore, as it can be transferred to all object-oriented metamodeling formalisms.

For the definition of rules specifying knowledge base used to infer evolution strategies, we have adopted an adequate formalism for logic programming Prolog [10]. Prolog is chosen because in one hand it is a language of knowledge representation [17] and in the other hand using inference rules eliminates programming to get eventual compound changes, the task is performed by the inference engine of Prolog. Furthermore, Prolog interpreters are developed in several languages, which facilitates the use of the prolog formalism.

The computation of the differences between metamodel versions is performed with The Eclipse plug-in EMF Compare [18]. This tool provides algorithms to calculate the delta between two versions of a model and visualizes them using tree representations. EMF Compare is capable of detecting the following types of atomic operations:

- Add: A model element only exists in the revised version.
- Delete: A model element only exists in the origin version.
- Update: A feature of a model element has a different value in the revised version than in the origin version.
- Move: A model element has a different container in the revised version than in the origin version.

4 Related Works

In this section we will give an overview of current metamodel and model co-evolution approaches and already implemented systems. Over the last few years, the problem of metamodel evolution and model co-evolution has been investigated by several works like [4], [7-9], [12-13], [19-21]. Currently, there are several approaches that focus on resolving inconsistencies occurring in models after metamodel evolution [3], a classification of these model migration approaches is proposed in [3]. This classification highlights three ways to identify needed model updates: manually,

based on operators, and by using metamodel matching. When manually approaches like in [19-21], updates are defined by hand. In operators based approaches, like [7],[13], metamodels changes are defined in terms of co-evolutionary operators [14]. Those operators define conjointly the evolution on the metamodel and its repercussion on the models. Finally, in metamodel matching, like [4], [9], [12], versions of metamodels are compared and differences between them are used to semi-automatically infer a transformation that expresses models updates. Manual specification approach like Flock [21] is very expressive, concise, and correctness is also assured but finds difficulties with large metamodels since there is no tool support for analyzing the changes between original and evolved metamodels [22]. Operator based approaches like [13] ensure expressiveness, automaticity, and reuse [23], it was been perceived as strong in correctness, conciseness and understandability [22] but its lack is in determining which sequence of operations will produce a correct migration. Analysis of existing model co-evolution approaches, and comparison results of some works [3], [24-25] has yielded guidance for defining some requirements to our approach. To take advantage of state-based and operator-based approaches, previously discussed. We have proposed an alternative solution where we applied a hybrid approach to define model migration. The solution presented in this paper has a number of similarities with the techniques illustrated in [13], but it differs from this approach because it takes as input results of a matching process. Therefore, it permits evolving models with different tools. Another, strength of our solution is the proposed reasoning mechanism, which allows finding different evolution strategies and consequently different migration strategies. Proposed solution minimizes as far as possible the user effort to migrate models. Thus user intervention is limited to a control task in the end of the process to validate results which permits to increase expressivity and correctness.

5 Conclusion

In this paper we have proposed an alternative solution to automate the co-evolution of models and metamodels. In our proposal we use a hybrid approach. It takes advantages from state-based and operator based approaches. This solution consists of using a library of coupled operation and also a knowledge base of changes definition

The benefits of this approach are numerous, notably automaticity of the co-evolution is augmented compared with other techniques because even for changes requiring specific information, we have predict automatic model migration with user assistance. Moreover, our solution is independent from any modeling environment. It is easily adapted to various modeling environment. Using an intelligent logic mechanism to infer compound changes and evolution strategies increase effectiveness of our proposal. This makes our solution distinguishable from existing works.

However, currently the evaluation of the proposed framework is not performed. For a complete validation, we will conduct case studies with industrial models. In the long term, we want to study the possibilities to extend our solution to support representation of semantic in models and preserving semantics within the migration process as introduced in [26].

References

1. Bézivin, J.: On the Unification Power of Models. Software and systems Modeling (SoSyM.) 4(2), 171–188 (2005)
2. Favre, J.M.: Meta-model and model co-evolution within the 3D software space. In: International Workshop on Evolution of Large-scale Industrial Software Applications ELISA 2003, Amsterdam, pp. 98–109 (2003)
3. Rose, L.M., Kolovos, D.S., Paige, R.F., Polack, F.A.C.: An analysis of approaches to model migration. In: Joint MoDSE-MCCM Workshop (2009)
4. Garcés, K., Jouault, F., Cointe, P., Bézivin, J.: Managing Model Adaptation by Precise Detection of Metamodel Changes. In: Paige, R.F., Hartman, A., Rensink, A. (eds.) ECMDA-FA 2009. LNCS, vol. 5562, pp. 34–49. Springer, Heidelberg (2009)
5. Amirat, A.: Contribution à l'élaboration d'architectures logicielles à hiérarchies multiples, Thèse de Doctorat en Informatique, Université de Nantes, France (2010)
6. OMG: MOF QVT Final Adopted Specification (2005), http://www.omg.org/docs/ptc/05-11-01.pdf
7. Wachsmuth, G.: Metamodel adaptation and model co-adaptation. In: Ernst, E. (ed.) ECOOP 2007. LNCS, vol. 4609, pp. 600–624. Springer, Heidelberg (2007)
8. Amirat, A., Menasria, A.: ne Gasmallah:Evolution Framework for Software Architecture using Graph Transformation Approach. In: The 12th International Arab Conference on Information Technology (ACIT'2011), Riyadh, Saudi Arabia, December 11-14, pp. 75–82 (2011)
9. Gruschko, B., Kolovos, D.S., Paige, R.F.: Towards synchronizing models with evolving metamodels. In: International Workshop on Model-Driven Software Evolution (2007)
10. Savoy, J.: Introduction à la programmation logique Prolog (2006), http://members.unine.ch/jacques.savoy/lectures/SemCL/Prolog.pdf
11. Didonet Del Fabro, M., Bézivin, J., Jouault, F., Breton, E., Gueltas, G.: AMW: A Generic Model Weaver. In: IDM 2005 Premières Journées sur l'Ingénierie Dirigée par les Modèles, Paris (2005)
12. Cicchetti, A.: Difference Representation and Conflict Management in Model-Driven Engineering, Phd thesis (2008)
13. Herrmannsdoerfer, M., Benz, S., Juergens, E.: Automatability of Coupled Evolution of Metamodels and Models in Practice. In: Czarnecki, K., Ober, I., Bruel, J.-M., Uhl, A., Völter, M. (eds.) MODELS 2008. LNCS, vol. 5301, pp. 645–659. Springer, Heidelberg (2008)
14. Herrmannsdoerfer, M., Vermolen, S.D., Wachsmuth, G.: An Extensive Catalog of Operators for the Coupled Evolution of Metamodels and Models. In: Malloy, B., Staab, S., van den Brand, M. (eds.) SLE 2010. LNCS, vol. 6563, pp. 163–182. Springer, Heidelberg (2011)
15. Jouault, F., Kurtev, I.: Transforming models with ATL. In: Model Transformations in Practice Workshop at MoDELS Montego Bay, Jamaica, pp. 128–138 (2005)
16. EMF Eclipse Modeling Framework, http://www.eclipse.org/emf
17. Gaizauskas, R., Humphreys, K.: XI A Simple Prolog-based Language for Cross-Classification and Inheritance. In: Proceedings of the 7th International Conference on Artifi-cial Intelligence: Methodology, Systems, Applications (AIMSA 1996), Sozopol, Bulgaria, pp. 86–95 (1996)
18. EMFCompare, Eclipse modeling Project, http://www.eclipse.org/emf/compare/

19. Sprinkle, J., Karsai, G.: A domain-specific visual language for domain model evolution. Journal of Visual Languages and Computing 15, 291–307 (2004)
20. Narayanan, A., Levendovszky, T., Balasubramanian, D., Karsai, G.: Automatic Domain Model Migration to Manage Metamodel Evolution. In: Schürr, A., Selic, B. (eds.) MODELS 2009. LNCS, vol. 5795, pp. 706–711. Springer, Heidelberg (2009)
21. Rose, L.M., Kolovos, D.S., Paige, R.F., Polack, F.A.C.: Model Migration with Epsilon Flock. In: Tratt, L., Gogolla, M. (eds.) ICMT 2010. LNCS, vol. 6142, pp. 184–198. Springer, Heidelberg (2010)
22. Rose, L.M., Herrmannsdoerfer, M., Mazanek, S., Gorp, P.V., Buchwald, S., Horn, T., Kalnina, E., Koch, A., Lano, K., Schätz, B., Wimmer, M.: Graph and model transformation tools for model migration. Software and System Modelling Journal (2012)
23. Herrmannsdoerfer, M.: COPE – A Workbench for the coupled evolution of metamodels and models. In: Malloy, B., Staab, S., van den Brand, M. (eds.) SLE 2010. LNCS, vol. 6563, pp. 286–295. Springer, Heidelberg (2011)
24. Iovino, L., Pierantonio, A., Malavolta, I.: On the Impact Significance of Metamodel Evolu-tion in MDE. Journal of Object Technology 11(3), 1–33 (2012)
25. Herrmannsdörfer, M., Wachsmuth, G.: Coupled Evolution of Software Metamodels and Models. In: Mens, T., Alexander, S., Cleve, A. (eds.) Evolving Software Systems, p. 404. Springer (2014)
26. Cicchetti, A., Ciccozzi, F.: Towards a Novel Model Versioning Approach based on the Sep-aration between Linguistic and Ontological Aspects. In: ME 2013 Models and Evolution Workshop, pp. 58–65 (2013)

Extracting and Modeling Design Defects Using Gradual Rules and UML Profile

Mohamed Maddeh[1(✉)] and Sarra Ayouni[2]

[1] SOIE, ISG Tunis, Le Bardo, Tunis, Tunisia
maddeh_mohamed@yahoo.com
[2] Faculty of Sciences of Tunis, Tunis, Tunisia
s_ayouni@yahoo.fr

Abstract. There is no general consensus on how to decide if a particular design violates a model quality. In fact, we find in literature some defects described textually, detecting these design defects is usually a difficult problem. Deciding which object suffer from one defect depends heavily on the interpretation of each analyst. Experts often need to minimize design defects in software systems to improve the design quality. In this paper we propose a design defect detection approach based on object oriented metrics. We generate, using gradual rules, detection rules for each design defect at model level. We aim to extract, for each design defects, the correlation of co-variation of object oriented metrics. They are then modeled in a standard way, using the proposed UML profile for design defect modeling. We experiment our approach on 16 design defects using 32 object oriented metrics.

Keywords: Object oriented metrics · Data Mining · Gradual rules · Design defects detection · UML profile

1 Introduction

Design defects which are also called design anomalies, refer to design situations that adversely affect the development of software like bad smells [9] and antipatterns [2]. The first one (i.e., bad smells) was proposed by Beck [9]. In fact, the author defines 22 sets of symptoms of common defects. The second one (i.e., anti-patterns) was introduced by Brown et al. [2]. A set of refactoring suggestions are associate for each defect type. Detecting these defects at the model level is a promising way to improve software maintenance process [4][6][21]. In addition, it is difficult to identify and express these anomalies as rules [17], since they are not formalized and based on a simple textual description.

In general, design defects are evaluated using rules in the form of metric/threshold combinations. Some works propose rules manually identified [1][17], other propose algorithms that generate these rules[5][11][14]. Both approaches are suffering from two major difficulties. The first one is due to the large number of possible metrics combinations, in fact, it is difficult to find the best suitable rule. The second problem is to find the best threshold for each metric. In this paper, we propose a predictive

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 574–583, 2015.
DOI: 10.1007/978-3-319-19578-0_47

design defects detection that focuses on model level in order to correct them before there propagation to the code. Also, instead of affecting a threshold for metrics, we generate, using gradual rules a correlation of co-variation of metrics characterizing the object oriented design defects. We model each defect using UML profile, defect are then represented as an UML class diagram summarizing the relevant information from the most significant textual descriptions in literature.

The remainder of the paper is structured as follows. In section 2, we present the related works. In section 3, we give the problem statement. In Section 4, we introduce the general process of the approach. In sections 5, we validate the proposed approach and section 6 is reserved for conclusion.

2 Related Works

Several studies have recently focused on detecting design defects in software using different techniques. In [14] authors propose a new framework M-RAFACTOR for the detection and correction of design defects based on object oriented metrics. Marinescu [9] defined a list of rules relying on metrics to detect what he calls design flaws of OO design at method, class and subsystem levels. Erni et al. [18] use metrics to evaluate frameworks with the goal of improving them. Another model refactoring is presented by Marc Van Kempen et al. [13], based on SAAT (Software Architecture Analysis Tool). It allows calculating metrics about UML models the metrics are then used to identify the flaws or anti-patterns. Authors represent the structure using class diagrams, and the behaviour of each class using statecharts. After that they examine the metrics for refactoring a centralized control structure into one that employs more delegation. For the four previous contributions it is difficult to manually define threshold values for metrics in the rules.Moha et al. [15], in their DÉCOR approach, they start by describing defect symptoms using an abstract rule language. These descriptions involve different notions, such as class roles and structures. In [11] defect detection is considered as an optimization problem. They propose an approach for the automatic detection of potential design defects in code. The detection is based on the notion that the more code deviates from good practices, the more likely it is bad.

3 Problem Statement

There are many open issues that need to be addressed when detecting design defects. In this paper, we first focus on how to define detection rules when dealing with quantitative information and then how to give a unified representation of defects specifications.

In fact, we notice that the textual description of design defects presented by authors depend on a subjective interpretations of analysts. As fact, for a same design we can find variable set of defects depending on the criteria's used by designer team. To bridge the gap between the description and the detection process, each design defect must be formalized for the standardization of the definition of symptoms detection. In this paper we intend to use gradual rules to formalize design defects. In the context of

our research the generated gradual rules are represented as a correlation of co-variation of object oriented metrics. Once, gradual rules identified each design defect is then modeled using the UML profile for design defects. We have proposed an UML profile for design defect modeling. It summarizes the most relevant information and replaces all textual descriptions existing in literature by one class diagram for each design defect.

4 The General Process

As presented in figure 1, we start with the domain analysis of the knowledge extracted from the textual description of design defects. In fact, domain analysis is a process in which information used in developing software systems is identified, captured, and organized to be reusable when creating new systems [8]. In our context, information about design defects must be well structured and reusable for the automated detection process. Thus, we have studied the textual descriptions of design defects. We present an antipattern example named the Blob.

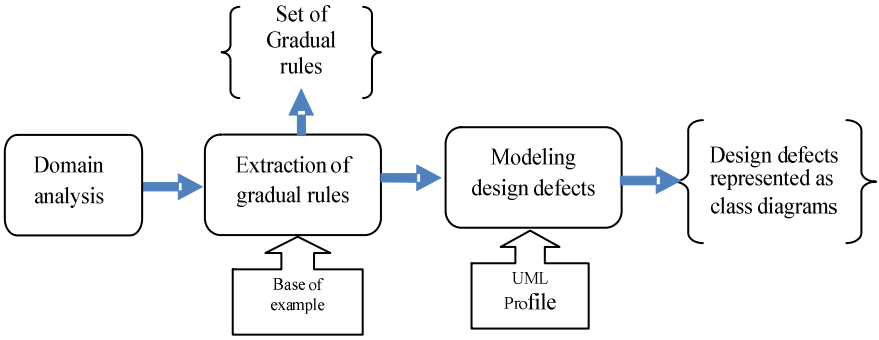


Fig. 1. General process

The Blob (called also God class [16]) corresponds to a large controller class that depends on data stored in surrounded data classes. A large class declares many fields and methods with a low cohesion. After the domain analysis for the Blob antipattern, we extract the relevant information. Indeed the blob is an interclass and behavioral defect, related to static and behavioral diagrams. The detection of the blob is based on the analysis of the class diagram and the sequence diagrams. As presented in table1, this research is based on 16 design defects.

These design defects are evaluated using object oriented metrics that are also identified at this step. Metrics must be measurable at model level, and useful for detection process. In our work we have identified 32 metrics. In what follows, we present some of these metrics:

Access To Foreign Data (ATFD) [12] represents the number of external classes from which a given class accesses attributes, directly or via accessor-methods.

Table 1. Classification of design defects

	Blob	SwissArmyKnife	Lava Flow	Poltergeists	FunctionalDecomposition	God Package	God Classes	Long Parameter List	Data Clumps	Divergent Change	ShotgunSurgery	Lazy Classes	FeatureEnvy	Comments	Data Classes	RefusedBequest
Structural					*			*	*							
Semantic														*		
Behavio-ral	*	*	*	*		*	*			*	*	*	*		*	*

Weighted Method Count (WMC) [3] is the sum of the complexity of all methods in a class.

Attribute Per method (APM) is defined as the ratio of the metrics Number of attributes (NOA) and (NOM).

After the metrics identification step we extract for each defect the most significant gradual rules that express the correlation of co-variation of the object oriented metrics. We propose an approach that uses knowledge from previously manually inspected projects, called defects examples.

4.1 Mining Gradual Rules

In our research, gradual rules are used to evaluate poor design by detecting bad smells and antipatterns. Mining gradual rule has been extensively used in fuzzy command systems. However, in last decade, the data mining community has been interested in extracting such kind of rules [7] [10] [19] [20]. Gradual rule convey knowledge of the form « the more/the less A, the more/the less B ». In our context, A and B are object oriented metric. We thus propose to extract rules such as « the more/the less Metrique1, the more/the less Metrique2…, the more/the less Metrique n », such that these metrics characterize a defect X. To the best of our knowledge, no previous study in the literature has paid attention to apply the extraction of gradual rules to the design defects detection. In the following section, we recall the key concepts of gradual rules mining.

Gradual Rules

We consider a data base defined on a schema containing m attributes (X1, …,Xm) defined on domains dom(Xi) provided with a total order. A data set D is a set of m-tuples of dom(X1),…,dom(Xm). In this scope, a gradual item is defined as a pair of an attribute and a variation {+,-}.The gradual item Xn+¸ means that the attribute Xn is increasing. It can be interpreted by the more A. A gradual itemset, or gradual tendency, is then defined as a non-empty set list of several gradual items.

For instance, the gradual itemset M =A+ B− is interpreted as, the more A and the less B. For example, the relation from Table 2 shows various items about disease symptoms.

Table 2. Disease symptoms

	Patients	Temperature	Lymphocyte	Hemoglobin
T1	P1	37.8	32	14
T2	P2	38.2	17	10
T3	P3	38.1	15	16

This table contains three tuples : {T1,T2,T3}, we study co-variations from one item to another one, as for example the variation of the temperature and hemoglobin. Too kinds of variations are considered: increasing variation and decreasing variation. Each item will hereafter be considered twice: once to evaluate its increasing strength, and once to evaluate its decreasing strength, using the + and - operators.

For example, let us consider the rule "The higher temperature and the higher hemoglobin then lower the lymphocyte" formalized by : R1= (Temperature + Hemoglobin + Lymphocyte -).

4.2 Mining Gradual Design Defect Rules

In this section, we present the extraction of gradual design defects rules. It is based on the GRITE algorithm [10], for GRadualITemset Extraction. For each design flaw, we identify the metric-based heuristics. The majority of works assign a threshold to each metric. The quality of the solution depends on the number of detected defects in comparison to the expected ones in the base of examples. The main limitation of this approach is that it is difficult to find the best threshold.

To overcome this problem, we present another type of correlation between object oriented metrics. To do so, we associate for each defect a metrics table; it represents the different metrics values for each occurrence (Oi) of all defects extracted manually from various projects (Pi). As example, we present in table 3 a part of the metrics table for the defect Data Class. The Data class defect creates classes that passively store data. Classes should contain data and methods to operate on that data.
Where, for a given class C we have:

PS: Package Size, NC: Number of Classes in the model, NOPM: Number Of Packages in the Model, NOC: Number Of Communications, is the number of messages sent by the class C, NMSC: Number Of Messages for the Same Class, is the number of internal messages from C to C, NCC: Number of Connected Classes, is the number of classes that communicates with the class C and NCM: Number of connected messages, is the number of messages sent to the class C.

The GRITE algorithm gives the most frequent sequences of metrics using the minsupport threshold. Where, the minsupport threshold aims at discovering subsets of items that occurs together at least a minsupport time in a database. If minsupport

Table 3. Data Class metrics

		ATFD	NOM	NOA	PS	NC	NOPM	NOC	NMSC	NCC	NCM
P1	O1	03	15	08	22	57	02	05	03	02	01
	O2	02	10	05	28	57	02	04	02	01	00
P2	O3	04	08	10	33	113	04	06	09	00	01
	O4	02	13	07	33	113	04	04	07	04	02
	O5	05	14	08	25	113	04	07	08	04	06
	O6	06	09	11	24	113	04	08	04	06	05
	O7	04	16	13	21	113	04	04	07	09	08
P3	O8	05	17	12	52	368	11	06	06	03	04
	O9	02	13	12	46	368	11	04	05	05	02

is set to be too large, no itemsets will be generated, if minsupport is set to be too small, huge number of itemsets will be generated. Fixing the minsupport threshold depend on the specificities of the problem.

In the context on design defect detection, almost we don't have a very large database comparing to other domains, that's why we set a minsupport value to be more than 0.5. It means that we will extract the gradual rules that occur at least in 50% of the transactions. We can decrease the minsupport thresholds if the program generates no rule, until having at least one rule.

4.3 Modeling Design Defects

Based on UML profile capabilities, we extend the UML metamodel to support and model all key concepts used for the specifications of design defects. We model each defect to create a catalogue of design flaws. We formalize a set of textual and informal design flaws description (avoiding any subjective interpretation) in a well-structured model enclosing all necessary information to deal with design defect detection.

Defined Stereotypes

In this section, we detail the defined stereotypes illustrated in figure 2: RefactoringIndicator is a super-class modeling all possible refactoring indicators. The design flows can be specialized as Antipattern, DesignPatternDefect, BadSmells.

Description contains a textual description of the design flaws. It represents the semantic aspect. The description stereotype is very helpful to understand the meaning of the design defect and the context in which it can be identified.

Metric represents the set of metrics useful forsoftware measurement and design flows detection. The measure of metrics is done over the static and/or dynamic UML diagrams. The UMLDiagram stereotype represents the UML diagrams attached to the metric concept. Each Design pattern defect is attached to a design pattern represented by the stereotype DesignPattern. RefactoringRepository indicates the name of the refactoring primitive, using the attribute PrimitiveName (For the design defects correction).

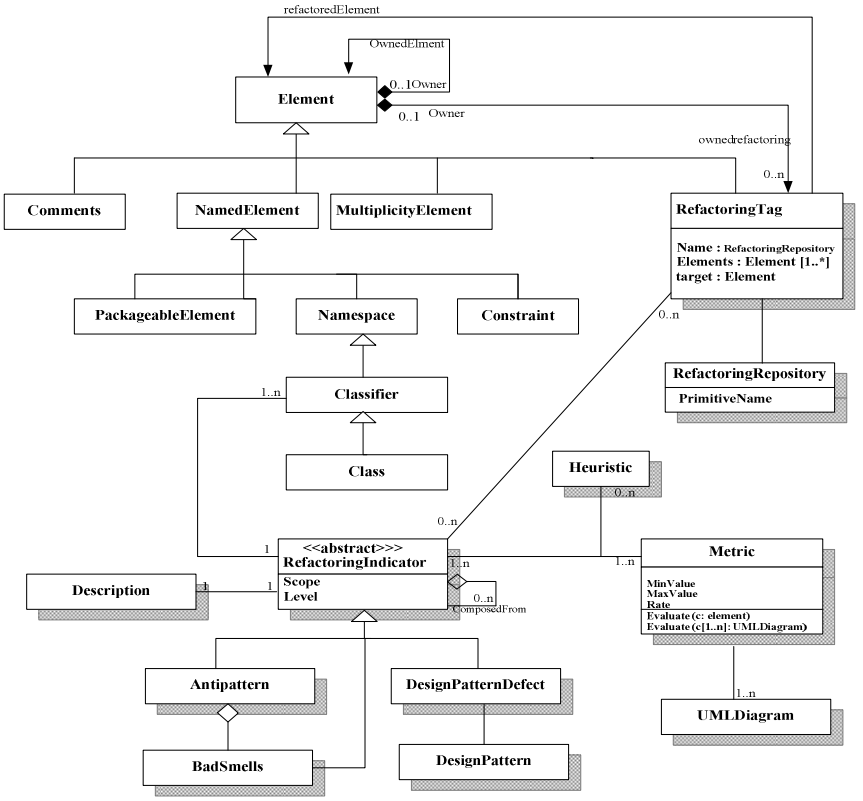


Fig. 2. The UML profile

Figure 3 in the next section presents the UML class diagram of the Data Class defect.

5 Experiments

The experimentations concern the defects Blob, Lazy class, Data class, Feature Envy and Lava flow, using three minsupport thresholds 0.5, 0.8 and 0.9. The Lazy class defect occurs when class isn't doing enough to pay for itself. Every additional class increases the complexity of a project. The Feature Envy defect occurs for methods that make extensive use of another class and may belong in another class. And the Lava flow defect represents the model elements that are not really used in the project due to an overestimation of needs.

The Table 4 presents our results. We choose to select gradual design defect rules that contain more than four metrics to guaranty significant results avoiding an overwhelming rule set. These results indicate the common conditions for the occurrence of a design defect. We have better rules for high minsupport value 0.8, because the

Table 4. Results

		Minsupport	
		0.5	**0.8**
Blob	***R1***	(ATFD+ PS+ NC+ NOPM-)	No Rule
	R2	(ATFD+ NOM+ NOA- NCM+)	
	R3	(ATFD+ NC+ NCM+ NCC+ NOM+)	
	R4	(ATFD+ NMSC+ NOA- PS+ NC+)	
Lazy class	***R1***	(NC+ NCC- ATFD- NCM- PS+)	(NCM- NOM – NC+ NCC- ATFD-) (NC+ NCC- ATFD- NCM- PS+)
	R2	(NCM- NOM – NC+ NCC- ATFD-)	
	R3	(ATFD- NC+ NCM- NOPM- NOM-)	
Data class	***R1***	(APM- ATFD- NC+ PS+ NCC-)	(NOM- PS+ NCC- NC+ NCM+)
	R2	(NOM- PS+ NCC- NC+ NCM+)	
FeatureEnvy	***R1***	(NIC+ NMSMC- NC+ PS+ NOPM-)	No Rule
Lava flow	***R1***	(NC+ NCC- ATFD- NCM- PS+)	(NCM- NOM – NC+ NCC- ATFD-)
	R2	(NCM- NOM – NC+ NCC- ATFD-)	(NC+ NCC- ATFD- NCM- PS+)
	R3	(ATFD- NC+ NCM- NOPM- NOM-)	(NIC- CM- APM- NC+ NOPM+)
	R4	(NIC- NMSMC- CM- NOM+)	
	R5	(NIC- CM- APM- NC+ NOPM+)	

rule is repeated at the majority of the defect occurrence (80%). We have lowest minsupport threshold 0.5 guaranty that the extracted rules occurs in at least 50% of the detected defects.

In our case, for a minsupport threshold equal to 0.9 we have no rules for all defects. We notice that the activity of design defects detection depends on the subjectivity of the designer. In fact, our research intends to help designer to improve the quality of models by offering a set of gradual rules characterizing the context in which could occur a design defect. All important information related to defects is now represented using the UML profile. In figure 3 we present an example for the Data Class defect.

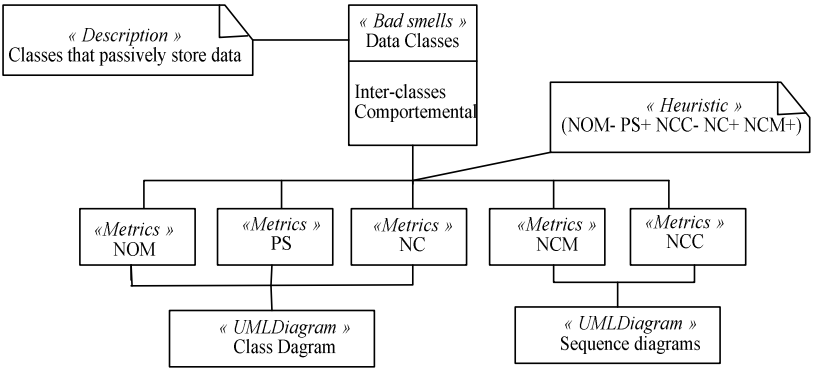


Fig. 3. Data Classes

6 Conclusion

Several design defect detection techniques have been proposed. Most of existing works relies on metrics rule-based detection, applied for the code level. However, it is difficult to identify and express these symptoms as rules [17], since they are not formalized. It is also difficult to find the best threshold for metrics. This work raised some interesting perspectives in order to detect design defects for model level based on the evaluation of correlation of metrics co-variation instead of threshold. We have also proposed an UML profile for design defect modeling. It fully supports design defects modeling needs. It allows antipatterns and bad smells modeling with one unified language. Using the UML profile for design defects, we unify software designer teams with a single and shared design defects specification.

References

1. Brito e Abreu, F., Melo, W.: Evaluating the impact of object-oriented design on software quality. In: The 3rd International SoftwareMetrics Symposium, pp. 90–99 (1996)
2. Brown, W.J., Malveau, R.C., McCormick, H.W.S., Mowbray, T.J.: AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis: Refactoring Software, Architecture and Projects in Crisis. John Wiley & Sons (1998b)
3. Chidamber, S., Kemerer, C.: A metrics suite for object oriented design. IEEE Transactions on Software Engineering 20(6), 476–493 (1994)
4. Corradini, A., Ehrig, H., Kreowski, H.-J., Rozenberg, G. (eds.): ICGT 2002. LNCS, vol. 2505. Springer, Heidelberg (2002)
5. Erni, K.: C, Applying design metrics to object-oriented frameworks. In: IEEE METRICS, pp. 64–74 (1996)
6. Hadar, E., Hadar, I.: The Composition Refactoring Triangle (CRT) Practical Toolkit: From Spaghetti to Lasagna. In: OOPSLA 2006, Portland, Oregon, USA. ACM (2006), 1-9593-491-X/06/0010
7. Hüllermeier, E.: Association rules for expressing gradual dependencies. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 200–211. Springer, Heidelberg (2002)
8. Frakes, W., Prieto-Diaz, R., Fox, C.: DARE: Domain Analysis and Reuse Environment. Annals of Software Engineering (5), 125–141 (1998)
9. Fowler, M., Beck, K., Brant, J., Opdyke, W., Roberts, D.: Refactoring: Improving the Design of Existing Code (1999)
10. Di-Jorio, L., Laurent, A., Teisseire, M.: Mining frequent gradual itemsets from large databases. In: Adams, N.M., Robardet, C., Siebes, A., Boulicaut, J.-F. (eds.) IDA 2009. LNCS, vol. 5772, pp. 297–308. Springer, Heidelberg (2009)
11. Kessentini, M., Kessentini, W., Sahraoui, H., Boukadoum, M., Ouni, A.: Design Defects Detection and Correction by Example. In: 19th IEEE International Conference on Program Comprehension (2011)
12. Marinescu: Detecting Design Flaws via Metrics in Object-Oriented Systems. In: Proceedings of TOOLS USA 2001, pp. 103–116. IEEE Computer Society (2001)
13. Van Kempen, M., Chaudron, M., Kourie, D., Boake, A.: Towards Proving Preservation of Behaviour of Refactoring of UML Models. In: Proceedings of SAICSIT 2005, p. 252 (2005)

14. Mohamed, M., Romdhani, M., Ghedira, K.: M-REFACTOR: A New Approach and Tool for Model Refactoring. ARPN Journal of Systems and Software (July 2011)
15. Moha, N., Guéhéneuc, Y.-G., Duchien, L., Meur, A.-F.L.: DECOR: A method for the specification and detection of code and design smells. Transactions on Software Engineering (TSE), 16 pages (2009)
16. Riel, A.J.: Object-Oriented Design Heuristics. Addison-Wesley (1996)
17. Marinescu, R.: Detection strategies: Metrics-based rules for detecting design flaws. In: Proceedings of the 20th International Conference on Software Maintenance, pp. 350–359. IEEE Computer Society Press (2004)
18. Marticorena, R., Crespo, Y.: Refactorizaciones de especializacion sobre el lenguaje modelo MOON. Technical Report DI-2003-02, Departamento de Informatica. Universidad de Valladolid (septiembre 2003)
19. Yahia, S.A.S.B.: Fuzzy set-based formalization of gradual patterns. In: SoCPaR, 2014, Tunis, Tunisia, pp. 434–439 (2014)
20. Ayouni, S., Laurent, A., Ben Yahia, S., Poncelet, P.: Fuzzy gradual patterns: What fuzzy modality for what result? In: Proceedings of the International Conference on Soft Computing and Pattern Recognition (SoCPaR 2010), Cergy, France (2010)
21. Zhang, J., Lin, Y., Gray, J.: Generic and Domain-Specific Model Refactoring using a Model Transformation Engine. In: Model-driven Software Development – Research and Practice in Software (2004, 2005)

An Approach to Integrating Aspects in Agile Development

Tadjer Houda[1,2(✉)] and Meslati Djamel[3]

[1] Computer Science Department
Badji Mokhtar University, Annaba, Algeria
[2] LabSTIC, Guelma University, POB 401
24000 Guelma, Algeria
tadjerh@yahoo.fr
[3] Computer Science Department, LISCO Laboratory,
Badji Mokhtar-Annaba University, Annaba, Algeria
meslati_djamel@yahoo.com

Abstract. Separation of concerns is an important principle that helps to improve reusability and simplify evolution. The crosscutting concerns like security, and many others, often exist before implementation, in both the analysis and design phases, it is therefore worthwhile to develop aspects oriented software development approaches to handle properly the concerns and ensure their separation.

Moreover agile methods attempt to reduce risk and maximize productivity by carrying out software development with short iterations while limiting the importance of secondary or temporary artifacts, however these approaches have problems dealing with the crosscutting nature of some stakeholders' requirements. The work presented in this paper aims at enriching the agile development using aspect oriented approaches. By taking into account the crosscutting nature of some stakeholders' requirements, the combination of the two approaches improves the software changeability during the repeated agile iterations.

Keywords: Aspect oriented · Constraints · Extreme programming · Separation of concerns · User stories

1 Introduction

Taking into account the concerns in the analysis phase is currently regarded as an important step that could have a positive impact on subsequent development phases and, consequently, there are several Aspect Oriented Requirement Engineering models (AORE models) such as MC AORE model [13], Quality AORE model [10], Vgraph model [16] and Theme/doc model [2].

Moreover agile methods have become, due to their pragmatism, favoured approaches for complex systems development. The use of the early separation of concerns in agile approaches is an important issue which can cause considerable fallout in terms of software development management and clear architectural structuring. We found in literature several agile approaches: Extreme Programming [8], Scrum [14], Feature-Driven

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 584–595, 2015.
DOI: 10.1007/978-3-319-19578-0_48

Development (FDD) [11], and Dynamic Systems Development Method (DSDM) [15], Crystal Methodologies [4], Adaptive Software Development (ASD) [5]. Combining aspects oriented approaches with these agile approaches, eliminates the tangling and the scattering of the code they produce and consequently reduces the effort of understanding and changing this code during the repeated agile iterations.

In this paper, we present a work which aims at combining the separation of concerns, requirements' engineering and the well known Extreme Programming approach (Xp), in order to achieve a synergy that enhances the Xp approach of development by a convenient handling of concerns. This work focuses on how aspect concepts can be integrated within the requirement level of the agile development context and particularly in the Xp approach.

In the rest of this paper, we present, briefly, the Xp approach, the aspect oriented requirements engineering. Thereafter, we explain the proposed combination. Then we describe some related work and, finally, we give a conclusion.

2 Agile Approaches

The agile approaches are a family of pragmatic development approaches built to deliver products on time, budget and with high quality. These approaches focus on strong customers' involvement and guarantee that they will be satisfied by their project. The Xp approach is one of the most commonly used among agile approaches [7]. The Xp approach provides a life cycle model of software which is used as a guide for organizing the development team. This model is presented in Figure1. User stories are an important concept in the Xp development and are usually the starting point of all the Xp processes. By choosing them, the customer kick starts the iteration process. They represent what the customer wants the system to do [8]. The system development is a succession of such iterations where the requirements are continuously being defined by means of user stories. These user stories should feed into the release-planning meeting and to the creation of the user acceptance tests.

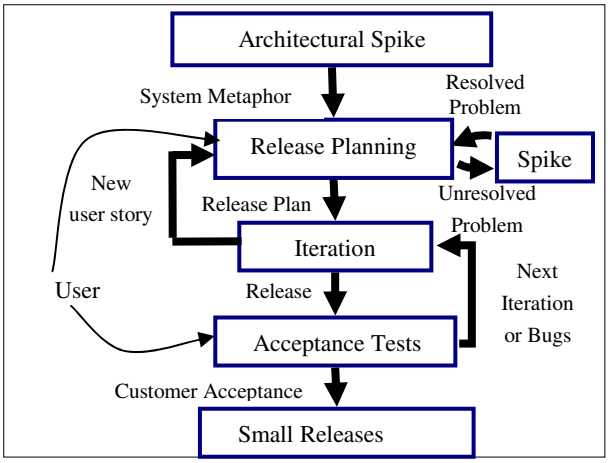


Fig. 1. Lifecycle of project Xp [7]

A release-planning meeting is used to create the release plan, which lays out the overall project. That is, the release plan indicates which user stories will be implemented and in which release this will happen. It also indicates how many iterations are planned and when each iteration will be delivered.

At the beginning of each iteration, an Iteration Planning meeting is held to determine exactly what will happen within that iteration. Such just-in-time planning is considered an easy way to stay on top of changing user requirements.

The acceptance tests are created from the user stories, written at the start of the project. Each iteration implements one or more user stories; these stories will be translated into a series of acceptance tests during the iteration.

3 Aspect Oriented Software Development (AOSD)

In object orientation, applications are modeled and implemented by decomposition of both the problem and solution space into objects, where each object embodies a single concern. However, some concerns still remain scattered throughout many different objects because they don't naturally fit within object boundaries. Such concerns (such as security, mobility, distribution, and logging) crosscut the other concerns. Aspect-oriented programming provides an elegant solution to this problem. Initially, the concept of aspect has been introduced in the context of programming. However, its use has become widespread to cover, among other things, analysis, design and evolution of applications.

To improve the software development, we need fully aspect oriented approaches that support aspects ranging from the analysis phase till implementation and testing. These approaches are often described as Aspect Oriented Software Development (AOSD) and encompass a range of techniques to achieve a better modularity.

Aspect-Oriented Software Development (AOSD) is regarded as a promising method that allows systematic identification, modularization, representation, and composition of such crosscutting concerns. Currently, it is commonly accepted that a good separation of concerns improves system modularity, reduces the complexity of software systems and the tangle of their code, facilitates reuse, improves comprehension, simplifies the integration of components and decreases the change, reducing the cost of adaptation, evolution and maintenance.

The requirement engineering is of vital importance because of its influence on the rest of the development. This is a starting point for many researches that aim at improving the separation of concerns at the requirements' level.

The motivation of these researches lies in reducing the cost of adaptation, evolution and maintenance. According to [3] the main activities in the requirements are: identification, capture, composition and analysis. Currently, several AORE models have been proposed and used as mentioned previously.

4 The Proposed Approach

Our work consists of integrating the aspect oriented approach in the Extreme programming approach and focuses particularly on the analysis phase. Our goal is to incorporate the concepts of AORE in the Xp process to make it more efficient and improve the productivity of the development team.

AORE's goal is the separation of concerns at the level of requirements, which may influence the way of using Xp in a project development. The agile development in general and Xp in particular can benefit from the AORE. In the Xp approach, the user stories represent the requirements of a system in the sense that the requirements are informally described by these stories from the customers. That is why our approach is mainly based on the user stories.

4.1 Steps of the Proposed Approach

Figure 2 summarizes the main steps we propose in our approach. We describe them shortly in the following.

Step 1: Identify User Stories and Constraints. In this stage customers and the development team meet to discuss the main functionalities to be achieved by the system. These functionalities are written by customers in the form of user stories on indexed cards. Each story has a short name that describes the functionality. Other details are added such as the risk to improve planning and performance of iteration. At the end of this stage, the customer must choose the stories to implement in the current iteration. Non functional requirements can address a variety of system needs and they can be considered as constraints on the system's behavior. Thus the customer must identify these constraints.

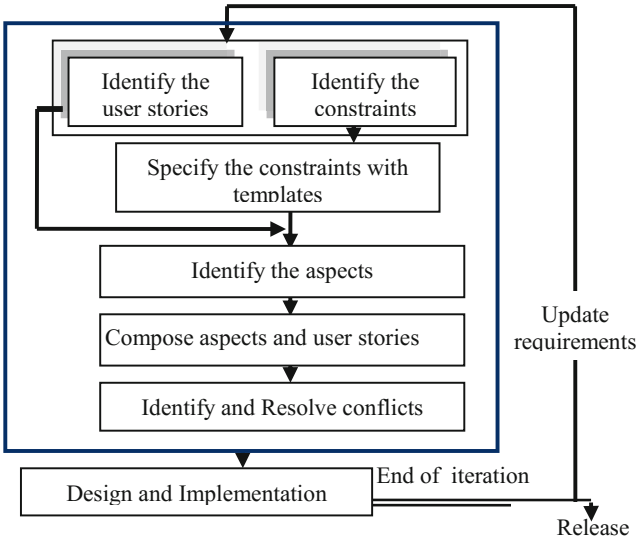


Fig. 2. Steps of the proposed approach

Step 2: Specify the Constraints. Based on the approach described in [10] the constraints are specified using templates as shown in table1.

Each constraint is defined as follows:

Table 1. Template for constraint

Constraint	Description
Name	Name of constraint.
Description	Description of constraint.
Influence	Lists of user stories affected by this constraint.
Priority	Priority assigned for this constraint.
Contribution	Represents how a constraint can be affected by other constraint. This contribution can be negative (-) or positive (+).

Step 3: Identify Aspects. If a constraint affects several user stories then this constraint is an aspect (taking in to account the information in row influence). In other words, if a constraint is triggered from several other user stories so it is considered as an aspect.

Step 4: Compose Aspects and User Stories. In this stage we try to compose crosscutting constraints and user stories in order to find the impact of an aspect on the requirements. We must define composition rules showing how an aspect influences behavior of a set of user stories.

Step 5: Identify and Resolve Conflicts. Identification of conflicts is based on the MCAORE technique [13] which uses a contribution matrix where each aspect may contribute negatively (-) or positively (+) to the others. If aspects have the same priority and contribute negatively then these aspects are in conflicts.

The conflicts in our approach are resolved through effective negotiation with customer who is part of the development team (on site customer).

Step 6: Design and Implementation. Xp like the other agile processes prioritizes pragmatic design for long-team change. The final set of user stories and aspects plus composition rules are used in the implementation, so an aspect oriented programming language could be used.

At the end of an iteration, the customer can check the product by the acceptance tests to detect errors or add other features. Following the change requirements, it is necessary to repeat the separation of concerns again.

4.2 Example

In this section we apply the previous steps to the creation of a website for the company SOUTH COAST NAUTICAL SUPPLIES to augment their print catalog. This example is taken from [5].

Step 1. Identify user stories and constraints. These are the user stories and constraints which are written by the stakeholders:

- User story 1: A user can do a basic simple search that searches for a word or phrase in both the author and title fields.
- User story 2: A user can put books into a "shopping cart" and buy them when he is done shopping.
- User story 3: To buy a book the user enters her billing address, the shipping address and credit card information.
- User story 4: A user can establish an account that remembers shipping and billing information.
- User story 5: A user must be properly authenticated before viewing reports.
- User story 6: An administrator can add new books to the site.
- User story 7: An administrator can delete a book.
- User story 8: An administrator can edit the information about an existing book.

Constraint C1: The system must support peak usage of up to 50 concurrent users.
Constraint C2: For audit purposes, all transactions in the system have to be kept.

In this iteration the first constraint shows that the system must support concurrent manipulation by at least 50 users which implies that there is a *multiple access* system. The second constraint is *audit* means that the action at each step is recorded.

By analysis of the stories identified for this iteration we can extract other constraints which are not written by the stakeholders. For example: User story 5, in this case the security must be guaranteed as the information provided by the user is personal data. The constraint identified here is *security*.

In the same way developers can also identify another constraint is the *login*.

Step 2. Specify the constraints.

Specification is as follows:

Table 2. Template for C1

Name	Multiple access
Description	Multiple users can use the system simultaneously.
Influence	multiple user stories
Priority	must have
Contribution	(+) Audit, (+) login

Table 3. Template for C2

Name	Audit
Description	The action at each step is captured and kept
Influence	multiple user stories
Priority	must have
Contribution	(+) Security, (+) multiple Access, (+) login

Table 4. Template for C3

Name	Security
Description	Only authorized users can access information.
Influence	multiple user stories
Priority	must have
Contribution	(+) Audit, (+) login

Table 5. Template for C4

Name	Login
Description	Provides the ability to connect and disconnect.
Influence	multiple user stories
Priority	must have
Contribution	(+) Audit, (+) Security, (+) multiple Access

Step 3. Identify aspects. From the table 2 to table 5, we deduce the following constraints:

Multiple access, A*udit, security and login* affect multiple user stories. So, these constraints are crosscutting and therefore represent aspects.

- Aspect1: *Multiple access*
- Aspect2: A*udit*
- Aspect3: *Security*
- Aspect4: *Login*

Step 4. Compose aspects and user stories. To combine aspects with the basic user stories we will first define composition rules indicating how these aspects influence the behavior of a set of basic user stories.

An example of composition rules in the case presented above is:

For safety, there is a recovery situation "Overlap", represented by the qualifiers before or after. Security for User story 5 is applied before viewing the report because we are in front of a protection in this case the composition rule will as follows:

- Security. Overlap. Before user story 5.

For the audit, there is also a situation of "Overlap", represented by before or after. The audit here is applied after each story affected recorded for each action, in this case the composition rule is as follows:

- Audit. Overlap. After all history

Step 5. Identify and Resolve conflicts. The contribution matrix which is symmetrical indicates whether aspects contribute positively or negatively. In our example aspects contribute positively and in this case no conflict appears in this iteration.

Table 6. The contribution matrix

Aspects / Aspects	Audit	Security	Multiple Access	Login
Audit		+	+	+
Security				+
Multiple Access				+
Login				

Step 6. Design and Implementation.

At the end of this iteration, suppose that the customer wants to add other constraints. A second iteration is then necessary. We describe it in what follows.

Step 1. Identify new user stories and constraints. These are user stories and constraints which are written by the customers:

- User story 9: A user can search for books by entering values in any combination of author, title and ISBN.
- User story 10: A user can view detailed information on a book. For example, number of pages, publication date and a brief description.
- User story 11: A user can put books into a "wish list" that is visible to other site visitors.
- User story 12: A user, especially a Non-Sailing Gift Buyer, can search for a wish list based on its owner's name and state.
- User story 13: A user can check the status of her recent orders.
- User story 14: If an order hasn't shipped, a user can add or remove books, change the shipping method, the delivery address and the credit card.
- User story 15: A user can view a history of all of his past orders.
- User story 16: A user can easily re-purchase items when viewing past orders.
- User story 17: A user can see what books we recommend on a variety of topics.
- User story 18: A user can remove books from her cart before completing an order.

Constraint C5: A customer must be able to find one book and complete an order in less than 90 seconds.

Step 2. Specify constraints.
For the next iteration, and due to changes in the requirements, a new specification is necessary (Table 7 to 11).

Table 7. Template for C1

Name	Multiple access
Description	Multiple users can use the system simultaneously.
Influence	multiple user stories
Priority	must have
Contribution	(+) Audit, (+) login, (-)Response time

Specification is as follows:

Table 8. Template for C2

Name	Audit
Description	The action at each step is captured and kept
Influence	multiple user stories
Priority	must have
Contribution	(+) Security, (+) multiple Access, (+) login, (-)Response time

Table 9. Template for C3

Name	Security
Description	Only authorized users can access information.
Influence	multiple user stories
Priority	must have
Contribution	(+) Audit, (+) login, (-)Response time

Table 10. Template for C4

Name	Login
Description	Provides the ability to connect and disconnect.
Influence	multiple user stories
Priority	must have
Contribution	(+)Audit, (+)Security, (+)multiple Access, (-)Response time

Table 11. Template for C5

Name	Response time
Description	Period of time in which the system responds to a service
Influence	multiple user stories
Priority	must have
Contribution	(-) Security, (-)multiple Access, (-) Audit, (-) login

Step 3. Identify aspects. From constraints which affect more than one user story, we deduce the following:

The first four (*Multiple access,* A*udit, security and login)* are already identified as aspects in the first iteration, the second constraint (response time) is crosscutting as they affect multiple user stories (taking into account the information in row influence) and therefore represent aspect. So for this iteration we add a new aspect:

- Aspect5: Response time

Step 4. Composed aspects and user stories. To combine the new set of aspects with the basic user stories we will define new composition rules indicating how these aspects influence the behavior of a set of user stories.

In this step we try to dial the new aspect 'response time' identified in the previous step with the stories it affects. As this aspect is required in parallel with the stories he forced this leads to use the relation "wrap".

For other aspects just add composition links with the stories of this iteration.

Step 5. Identify and Resolve conflicts. In our example the contributions between aspects are presented in the following table:

Table 12. The contribution matrix

Aspects / Aspects	Audit	Security	Multiple Access	Login	Response time
Audit		+	+	+	-
Security				+	-
Multiple Access				+	-
Login					-
Response time					

This table indicates the presence of conflicts between some aspects, if these aspects apply to the same user stories with the same priority. For example security and response time contribute negatively to each other. They constrain each other's behavior and have the same priority and apply to the the same user story, thus a conflict is arise.

As in this iteration we are faced with a conflict, it is resolved by negotiations between the customer and the development team.

Step 6. Design and Implementation.

The iterations are repeated until there is no need to add or update requirements.

5 Related Work

Several approaches are intended to identify crosscutting concerns during the early stages of development [9]. Aspect Oriented Requirement Engineering (AORE), described in [13], proposes a model for Aspect Oriented Requirement Engineering that supports the separation of crosscutting properties at the requirements level. Concerns and associated requirements are identified from different viewpoints. The rules of composition are defined using XML. In [10], functional requirements are specified using use cases based approach. The quality attributes are detailed extensively in a template, which among other details also lists down the

decomposition of the quality attribute, priorities (max, high, low, and min), and influence of the quality attribute. By observing the influence of a quality attribute and associated requirements, crosscutting concerns (quality attributes) are identified. Here, a set of UML models are integrated to the crosscutting quality attributes. Baniassad and Clarke [2] propose the Theme approach that does not identify the crosscutting concerns from traditional requirements engineering approaches. They introduced the concept of action view, clipped action view, base themes, and crosscutting themes to provide support for the aspect orientation in the analysis and design. Theme supports activities of requirements analyses. The results of analyses are mapped to the UML models. Despite the diversity of these approaches, no one among them takes into account the agile development. the FDD approach takes into consideration the integration of aspects in a contextual agile development [12]. In [1], a method is proposed for unifying agile and AO requirements analysis approaches.

6 Conclusion

The Xp approach is a development approach that can produce quickly software of high quality. This development approach may benefit from aspect-oriented requirements engineering approaches in a variety of ways.

The work presented in this article is a proposition of integration between separation of concerns and requirements engineering in an agile development context and particularly in the extreme programming approach. The main contribution of our work is its focus on the user stories and constraints as the starting point of the integration. Our approach is still at its beginning and we are now using it for more complex systems.

References

1. Araujo, J., Ribeiro, J.C.: A scenario and aspect-oriented requirements agile approach. International Journal of Computing Science and Applications 5(3b), 69–92 (2008)
2. Baniassad, E., Clarke, S.: Aspect-oriented analysis and design: Theme Approach. Addison Wesley professional (2005)
3. Baniassad, E., Clements, P.C., Araujo, J., Moreira, A., Rashid, A., Tekierdogan, B.: Discovering Early aspects. IEEE Software (January/February 2006)
4. Cockburn, A.: Crystal methodologies: The Cooperative Game. Addison-Wesley (2006)
5. Cohn.M, User Stories Applied: For Agile Software Development, Addison Wesley (March 2004)
6. Highsmith.J.: Adaptive Software Development. Dorset House, New York (2000)
7. Hunt.J, Agile software construction (2006)
8. Kent, B.: Extreme Programming Explained: Embrace Change. Addison Wesley (2000)
9. Moreira, A., Araújo, J.: The Need for Early Aspects. In: Fernandes, J.M., Lämmel, R., Visser, J., Saraiva, J. (eds.) Generative and Transformational Techniques in Software Engineering III. LNCS, vol. 6491, pp. 386–407. Springer, Heidelberg (2011)

10. Moreria, A., Araújo, J., Brito, I.: Crosscutting quality attributes for requirements engineering. In: SEKE2002: Fourteenth International Conference on Software Engineering and Knowledge Engineering, Ischia, Italy, July 15-19 (2002)
11. Palmer, S.R., Felsing, J.M.: A Practical Guide to Feature-Driven Development. Addison-Wesley (2002)
12. Pang, J., Blair, L.: Refining feature driven development - a methodology for early aspects. In: Early Aspects: Aspect-Oriented Requirements Engineering and Architecture Design (2004)
13. Rashid, A., Moreira, A., Araújo, J.: Modularization and composition of Aspectual Requirements. In: 2nd International Conference on Aspect- Oriented Software Development, Boston, pp. 11–20 (2003)
14. Schwaber, K., Beedle, M.: Scrum: Agile Software Development. Prentice-Hall (2002)
15. Stapleton.J, Dynamic Systems Development Method:The method in practice. Addison-Wesley (1997)
16. Yu, Y., Leite, J.C.S., Mylopoulos, J.: From goals to aspects: discovering Aspects from requirements goal models. In: The 12th IEEE International Requirements Engineering Conference, Kyoto, Japan (2004)

Software Engineering: Checking and Verification

On the Optimum Checkpointing Interval Selection for Variable Size Checkpoint Dumps

Samy Sadi(✉) and Belabbas Yagoubi

University of Oran1 Ahmed Benbella,
Department of Computer Science
Oran, Algeria
{samy.sadi.contact,byagoubi}@gmail.com

Abstract. Checkpointing is a technique that is often employed for granting fault tolerance for applications executing in failure-prone environments. It consists on regularly saving the application's state in another and fault independent storage such that if the application fails, it can be continued without necessarily restarting it. In this context, fixing the checkpointing frequency is an important topic which we address in this paper. We particularly address this issue considering hybrid fault tolerance and variable size checkpoint dumps. We then evaluate our solution and compare it with state of the art models, and show that our solution brings better results.

Keywords: Optimum Checkpointing Interval · Hybrid Fault Tolerance · Variable Size · Simulation

1 Introduction

Since the accession to information technologies, a lot of efforts have been devoted by the research community in order to make computing systems more fault-tolerant and more reliable. Initially, reliability was chiefly sought to avoid job resubmissions and to lower resource utilization, in a context where the job average length is ever-growing especially after the emergence of computational science and high-performance computing (HPC). Afterwards, and notably due to the advent of Cloud Computing and due to the increasing number of business-sensitive applications, a new practical and financial dimension appeared which drew even more attention to reliability.

The fact is that computing systems are failure-prone and failures are getting more frequent as new systems appear. The reason behind this is not because their components are getting less reliable. Actually, future hardware components and in particular newer generation chips are expected to keep failure rates similar to those of the current generation [14]. However, the number of components per any single system has considerably increased since the last few years and is continually increasing which led to a lower overall system mean time between failures (MTBF). So as to emphasize these lines, a study [2] on large-scale HPC systems has observed an MTBF in the 6.5h–40h range. This value when extrapolated for a peta-scale system, corresponds to a MTBF of only 1.25 hours [11]. Latterly,

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 599–610, 2015.
DOI: 10.1007/978-3-319-19578-0_49

another study [16] has observed that in a typical Cloud datacenter, a proportion of 8% of the machines can expect to see at least one failure each year.

In this context, checkpoint-restart or checkpointing has been developed in order to leverage fault tolerance in computing systems. This technique consists on taking frequent snapshots of any job's state, and on saving it on a secondary and fault-independent machine [5]. When the job fails on its primary machine, the saved state on the secondary machine is used to restore the job's state and to continue it. Thus, the job does not need to restart from scratch and a lower execution delay can be expected in failure-prone environments.

A central concern when implementing such fault tolerance technique is about selecting the frequency of checkpointing or the delay between taking two checkpoints. A high checkpointing frequency will ensure to have at any moment a very recent snapshot of the job's state, thus minimizing the potential rework delay after a failure. But in the same time, this will induce a significant overhead to the job execution if the occurrence of failures is very low or nonexistent. In another hand, a too low checkpointing frequency is obviously not a wise choice either, as this will lead to a noteworthy rework delay after failures.

In the present study, the issue under scrutiny is precisely about determining the best checkpointing frequency, also known as the checkpointing interval selection problem. We particularly address this problem considering a variable checkpointing overhead. We assume that the checkpointing overhead, or the time which is necessary to save a job's state into an external device is a function of the previous computing phase's delay. Besides, as new research has been undertaken for hybrid fault tolerance and in particular to predict the occurrence of failures with fair results [7,12,13], we also take into consideration this aspect in order to reduce checkpointing frequency.

The organization of this paper is as follows. In the next section, we discuss related work. In section 3, we define and formalize the problem and we present the brought solution. In the penultimate section, we evaluate our solution and we discuss the results. Finally, in section 5 we conclude and we give an overview of our future work.

2 Related Work

Checkpointing and in particular checkpointing interval selection has been extensively studied in the past. In this section, we give a chronological review of most prominent research efforts in the literature.

One of the first contributions was made by Young [17] who managed to give a first order approximation to the optimum checkpointing interval. Young considered that the overhead which is due to checkpointing is (1) constant and independent from the computing phase, and (2) is negligible when compared to the system MTBF. Furthermore, failures occurrence was assumed to be independent and exponentially distributed following a known MTBF. This last assumption, even if adopted in many contributions [4,6,8,15,17,18], is only true for the first occurrence of the failures. In fact, failures cluster in time and a failure is more likely to happen after a first failure [16].

In [6], the author showed that the optimum checkpointing interval is deterministic and is a function of the system load. In particular, the author proposed a queuing model where the duration of service interruptions is directly computable knowing the past history of the system.

An $O(n^3)$ algorithm has been proposed in [15] to select the $(n-1)$ potential checkpoint locations which can minimize a given job's execution time. To do so, the authors considered that a job consists of a set of tasks and that checkpoints can only be taken between two consecutive tasks (and not during the task's execution). As opposed to the so far presented contributions, in this contribution authors assumed that checkpointing overhead is not constant and is task-dependant.

An aperiodic checkpointing approach where the checkpointing interval is varying from one checkpoint to another has been proposed in [9]. The authors considered a general failure-rate and no assumption was made regarding the distribution of the failures. Nevertheless, this research specifies that when the distribution of failures is exponential, the optimum placement of checkpoints is equidistant.

A higher order estimate of the optimum checkpointing interval has been provided in [4]. Daly has undertaken to continue the groundwork initiated by Young in [17], and kept most of his assumptions particularly regarding the checkpointing overhead and the failures distribution. However, Daly generalized Young's solution considering the case where the checkpointing overhead is not negligible compared to the system's MTBF.

In [10], the authors proposed an approach for checkpoint placement under incomplete failures information and when the failures distribution is unknown. The min-max principle has been employed to this extent.

An hybrid fault tolerance approach has been proposed in [8]. In this approach, it is assumed that the system has fault-prediction capabilities [7,12,13] which can be used to reduce checkpointing frequency. The authors proposed to partition the job's execution using a preset time interval. At each interval, a decision stage takes place where (1) the job is checkpointed, (2) the job is migrated or (3) no action is taken.

The checkpointing scheduling complexity has been analyzed in [3]. In this research, no assumption was made regarding failures distribution, and checkpointing overhead was assumed to be variable. The authors stated that the checkpointing problem is NP-hard even in the simple case where the failures distribution is uniform. In addition, a dynamic programming algorithm has been proposed to solve the problem.

In [18], the authors exploited failures prediction capabilities in order to define a new formula for computing checkpointing interval. Two main metrics were considered, namely the precision and the recall of failures prediction. These two metrics respectively characterize the capacity of the system to not make false predictions of failures, and the capacity of the system to predict all future failures. We also consider these two metrics in current paper, but we consider variable checkpointing overhead.

3 The Checkpointing Interval Model

In this section, we develop our model for estimating the optimum checkpointing interval considering a variable checkpointing overhead. We draw our inspiration in the model proposed by Young [17] and later used in many other research [4,18].

In the next lines, we first describe the checkpointing process in both the situations where a failures prediction mechanism is employed or not. Then, we give the cost function we want to optimize. Next, and before solving the equation we place some assumptions regarding the failures distribution and the checkpointing overhead function. Once the assumptions placed, we solve the cost function and the optimum checkpointing interval is quantified. Finally, we address a special issue as regards to if the checkpointing overhead function is bounded.

The main symbols used in this paper are described in table 1.

Table 1. Description of used symbols

Symbol	Description
t	Checkpointing interval.
$\delta(t)$	Checkpointing overhead, or the length of the save phase considering t is the length of the compute phase.
δ_{max}	The maximum length of the checkpointing overhead.
α and β	Values characterizing the job's checkpointing overhead.
R	Restart delay before a job can continue on a secondary machine.
C	Total number of real failures during job execution.
C_{tp}	Total number of failures that were predicted (true positives).
C_{fp}	Total number of failures that were predicted but will actually not happen (false positives).
C_{fn}	Total number of failures that were not predicted (false negatives).
p	Precision value in the $[0, 1]$ range for failures prediction.
r	Recall value in the $[0, 1]$ range for failures prediction.
$n_i(t)$	Number of successful computing phases between the $(i-1)^{th}$ and the i^{th} failure event.
$w_i(t)$	Last computing phase's delay just before the i^{th} failure event happens.
$l_i(t)$	Time lost due to the i^{th} failure event.
$L(t)$	Time lost due to checkpointing and considering all failure events.
S	The submitted job length, or the execution time needed by the job to complete.
M	The mean time between failures (MTBF).
t_{opt}	Optimum Checkpointing interval.

3.1 The Fundamental Checkpointing Process

The checkpointing process consists of multiple sequences of a computing phase where the job is normally executing followed by a save phase where the job's state is written to an external storage. During each save phase the job execution is paused and the job continues in the next computing phase.

The checkpointing interval t represents the length of the computing phase which is also the delay between two consecutive save phases. In this paper, we assume that the save phases are equidistant and that the checkpointing interval stays unchanged during all the job's execution.

During a job's execution one or more failures may happen. After each failure, a restart delay R is necessary for the job to be continued. Moreover, there is an added rework delay $w_i(t)$ corresponding to the not saved job progress due to the i^{th} failure. Refer to Fig.1 for an overview of a job's execution in a context where checkpointing is used.

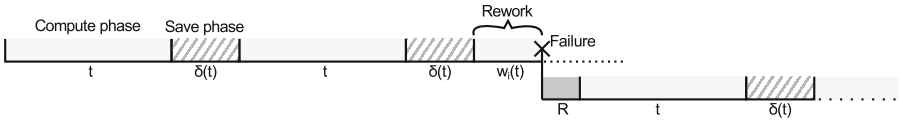


Fig. 1. The Checkpointing Process

3.2 The Hybrid Checkpointing Process

As previously discussed, there is a rapidly growing literature on failures prediction techniques. These techniques can be employed on top of periodic checkpointing in order to trigger additional save phases when a failure is predicted. After the save phase, the job is immediately continued on the secondary machine and no rework delay is necessary as the job state on the secondary machine is up to date. Another important feature of such process, is that checkpointing frequency can be reduced. The hybrid checkpointing process is displayed in Fig.2.

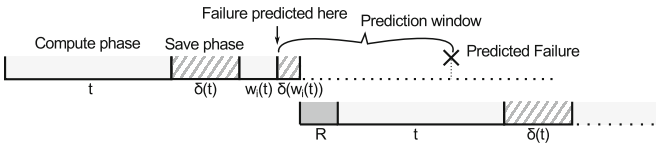


Fig. 2. The Hybrid Checkpointing Process

The level of trust of failures prediction is characterized by three different metrics, namely: the precision(p), the recall(r) and the prediction window.

The precision is the fraction of true positive predictions (C_{tp}) that are made by the system when compared to all the predictions made. In fact, the system might do an erroneous prediction of a future failure. These erroneous predictions are also designated as false positives (C_{fp}).

$$p = \frac{C_{tp}}{C_{tp} + C_{fp}} \tag{1}$$

The recall value represents the ability of the system to predict all future failures. It is the fraction of true positive predictions on the total number of failures. The missed failure predictions are designated as false negatives (C_{fn}) and is used when computing the recall:

$$r = \frac{C_{tp}}{C_{tp} + C_{fn}} \tag{2}$$

The prediction window is the time interval which is left before a predicted failure happens. A good prediction window is big enough so that a save phase can be engaged and completed before the failure happens. If the prediction window is too small, then the failure prediction is pointless as the save phase will not complete. Such predictions are consider to be false negatives in the current paper.

3.3 Cost Function

We define our cost function by considering the time lost due to checkpointing in different scenarios. We start by considering the time lost due to at most one single failure event. After that, we generalize the results to more than one failure event. A failure event is either an actual failure happening, or a failure prediction event.

Single Failure Event Cost Function. We identify four cases as regards to the time lost in the checkpointing process when compared to a job executing in a regular and failure immune system. We present the cost functions for each case. We assume in those that $n_i(t)$ is the number of successful computing phases between the $(i-1)^{th}$ and the i^{th} failure event.

The first case is when no failure events happen and the job executes normally on the primary machine. In this regard, the time lost is the sum of the delays spent in all the save phases.

$$l_{i,1}(t) = n_i(t) \cdot \delta(t) \tag{3}$$

The second case is when an unpredicted failure happens. Hereof, the time lost consists of all the save phases plus a restart delay (R) and eventually some rework time ($w_i(t)$) which has not yet been saved.

$$l_{i,2}(t) = n_i(t) \cdot \delta(t) + R + w_i(t) \tag{4}$$

The third case is when the system predicts the occurrence of a failure that will really happen (ie: a true positive). In this case, the time lost consists of all the save phases plus a restart delay. The last save phase being the one initiated after the failure prediction. Besides, no rework time is induced in this situation.

$$l_{i,3}(t) = n_i(t) \cdot \delta(t) + R + \delta(w_i(t)) \tag{5}$$

The fourth and final case is when the system predicts a failure which will not happen (ie: a false positive). The time lost is the same as in the third case.

$$l_{i,4}(t) = l_{i,3}(t) = n_i(t) \cdot \delta(t) + R + \delta(w_i(t)) \tag{6}$$

Final Cost Function. We can now express the lost time in the general case and considering multiple failure events. Note that we exclude the special case where $C = 0$, as in this case, the optimum checkpointing interval is obvious and is infinity. Furthermore, we consider that the system where the job runs is failure prone and $C > 0$. The total cost function is thus as follows:

$$L(t) = \sum_{i=1}^{C_{fn}} l_{i,2}(t) + \sum_{i=1}^{C_{tp}} l_{i,3}(t) + \sum_{i=1}^{C_{fp}} l_{i,4}(t) \tag{7}$$

Considering that the job length is S, we can estimate the number of successful computing phases for the job to complete as follows:

$$N(t) = \frac{S}{t} = \sum_{i=1}^{C} n_i(t) \tag{8}$$

The equation 7 can be expanded as follows.

$$\begin{aligned} L(t) = S\,\frac{\delta(t)}{t} + (C + C_{fp})\,R + C_{fn}\,w_i(t) + \\ (C_{tp} + C_{fp})\,\delta(w_i(t)) + \delta(t) \sum_{i=1}^{C_{fp}} n_i(t) \end{aligned} \tag{9}$$

The optimum checkpointing interval t_{opt} is the value of t which produces the smallest time lost. In other words, it is the minimum of the function $L(t)$.

3.4 Solving Assumptions

Assumption on the Checkpointing Overhead. We assume that the dump size produced by jobs is linear and is a function of time. Thus, and considering a fixed bandwidth allocation, the checkpointing overhead is function of t and is linear. It can be expressed as follows:

$$\delta(t) = \alpha \cdot t + \beta \tag{10}$$

Of course, this simplistic formulation may not fit for any type of job. One main issue, is that the job's dump size may be bounded with a maximum. Another issue, is that the job's dump size may not be linear.

We will address the first issue in section 3.6. For the second issue, we consider that a fairer approximation can be made using the previous formula as when compared to constant checkpointing overheads. But we do not further address this issue in this paper.

First Assumption on the Failures Distribution. We assume that failures are independent and follow an exponential distribution of mean M. The literature have shown that this is usually only true for the first occurrence of the

failure on the machine and tends to be false once the machine have been repaired. Thus this is only pertinent if we did consider repairs.

As the failures distribution is known, we can now estimate the number of failures that will happen for a job of length S. As previously stated, we need $N(t)$ successful computing phases to complete the job. Besides, we know that the probability to complete one computing phase (followed by a save phase) is:

$$Q_1(t) = P(x > t+\delta(t)) = 1 - P(x \leq t+\delta(t)) = 1 - (1 - e^{-(t+\delta(t))/M}) = e^{-(t+\delta(t))/M} \tag{11}$$

Thus the number of tries to complete $N(t)$ computing phases is:

$$Q_n(t) = \frac{N(t)}{P(x > t + \delta(t))} = N(t) \cdot e^{\frac{t+\delta(t)}{M}} \tag{12}$$

Finally, the number of failures is:

$$C = Q_n(t) - N(t) = N(t) \cdot (e^{\frac{t+\delta(t)}{M}} - 1) \tag{13}$$

Second Assumption on the Failures Distribution. We assume that the mean time between failures (M) in the system is big enough such that the checkpointing interval (t) and the restart delay (R) are negligible when compared to it. And as a direct consequence to this assumption, the checkpointing overhead ($\delta(t)$) is also negligible when compared to M.

We know that the first degree Taylor's series expansion is a good approximation for small values. Thus, the equation 13 can be reformulated as follows:

$$C = N(t) \cdot \frac{t + \delta(t)}{M} = S \cdot \frac{t + \delta(t)}{t \cdot M} \tag{14}$$

Assumption on the Rework Time. We assume that on average, the failure event happens in the middle stage as regards to the computing phase. Because, failures are independent and exponentially distributed, this value is fair enough.

$$w_i(t) = \frac{t}{2} \tag{15}$$

Assumption on the Moment of Failures Predictions. We need a final assumption as regards to the moment of the failures predictions when those are false predictions (false positives). We assume that those are uniformly distributed as regards to the whole job execution. In other words, we can make the following approximation:

$$\sum_{i=1}^{C_{fp}} n_i(t) = N(t) \cdot \frac{C_{fp}}{C} = \frac{S}{t} \cdot \frac{C_{fp}}{C} \tag{16}$$

3.5 Solution

Before relaxing the previous assumptions, it is worth to note that C_{tp}, C_{fp} and C_{fn} can be expressed as follows (based on equations 1 and 2):

$$C_{tp} = r \cdot C \tag{17}$$

$$C_{fp} = r \cdot C \cdot \frac{1-p}{p} \tag{18}$$

$$C_{fn} = (1-r) \cdot C \tag{19}$$

Now, after relaxing the assumptions on 9 we obtain (assuming K is a t independent variable):

$$L(t) = \frac{S\,(\alpha+1)(\alpha r - p\,r + p)}{2\,p\,M}\,t + \frac{S\,\beta\,(r\,\beta + (r - r\,p + p)\,(R+M))}{p\,M}\,t^{-1} + K \tag{20}$$

The optimum checkpointing interval is the minimum of the function $L(t)$. We also know that $L(t)$ attains its minimum when its first derivative function is zero. We thus need to solve:

$$\frac{\mathrm{d}}{\mathrm{d}t}L(t) = 0 \tag{21}$$

From equation 20 we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}L(t) = -\frac{S\,\beta\,(r\,\beta + (r - r\,p + p)\,(M+R))}{p\,M}\,t^{-2} + \frac{S\,(\alpha+1)(r\,\alpha - r\,p + p)}{2\,p\,M} \tag{22}$$

We can thus compute optimum checkpointing interval t_{opt} as follows:

$$t_{opt} = \sqrt{\frac{2\,\beta\,((M+R)\,p - (M+R)\,p\,r + (M+R+\beta)\,r)}{(\alpha+1)\,(p - p\,r + \alpha\,r)}} \tag{23}$$

We can get rid of the R and the β terms since we assumed that the restart delay and checkpointing overhead ($\delta(t) \Rightarrow \beta$) are negligible when compared to M. Thus, we obtain after simplification the final formula for the optimum checkpointing interval:

$$t_{opt} = \sqrt{\frac{2\,\beta\,M\,(p - p\,r + r)}{(\alpha+1)\,(p - p\,r + \alpha\,r)}} \tag{24}$$

We can note that the restart delay does not appear in the previous function which agrees with the result brought by Daly [4]. Besides when $r = 0$ and $\alpha = 0$, in other words when no failures predictions are made and when constant checkpointing overhead is assumed, then the optimum checkpointing interval is the same as the one predicted by Young [17]. However, the formula 24 is slightly different from the result brought in [18] when assuming $\alpha = 0$.

3.6 Bounding the Checkpointing Overhead

In the following lines, we address the issue related to when the checkpointing overhead is bounded with a known maximum (δ_{max}). Such use case can be envisioned if the job works on fixed size files. Once a file is totally modified the checkpoint dump size will be the same even if further modifications are made on that file. Thus we can write the following equation:

$$\delta(t) \leq \delta_{max} \tag{25}$$

Replacing $\delta(t)$ using equation 10, t can be bounded and the new optimum checkpointing interval is as follows:

$$t_{opt}^{'} = \min\left(t_{opt}, \frac{\delta_{max} - \beta}{\alpha}\right) \tag{26}$$

4 Evaluation

We have used the ACS simulator [1] in order to simulate and evaluate our model against state of the art models.

We have run multiple simulations for each tested model using the following simulation input. First, we consider different failure properties. Therefore, different system MTBF values are tested ranging from 1 hour to 10^4 hours. As regards to the precision and recall, we have tested two combinations corresponding to the results reported in the literature [7]. Next, for each simulation, a total of 1000 jobs with a mean length of 500 hours are launched. And finally, for each job, we have considered empirical values for the α, β and R parameters.

Given a simulation input, the simulator computes the finish time for each job and the average job finish time considering all the jobs. This value is used to compare different models including Young's [17], Daly's [4] and Zhu's [18]. We have observed in different simulation scenarios that when using our model, we obtain lower average job finish time when compared to other models. For brevity, we only include the results of the comparison of our model with Zhu's [18] model

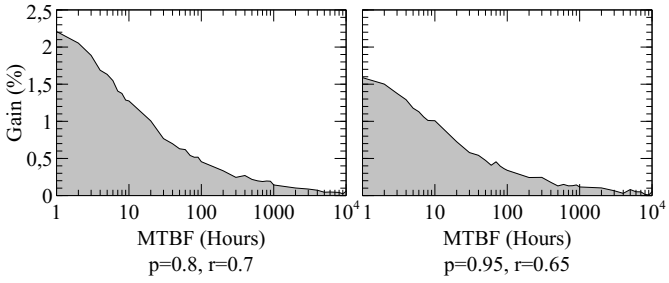


Fig. 3. Time gain percentage of Our Model when compared to Zhu's Model [18] assuming a Constant Checkpointing Overhead ($\alpha = 0$, $\beta = 5min$, $R = 10min$)

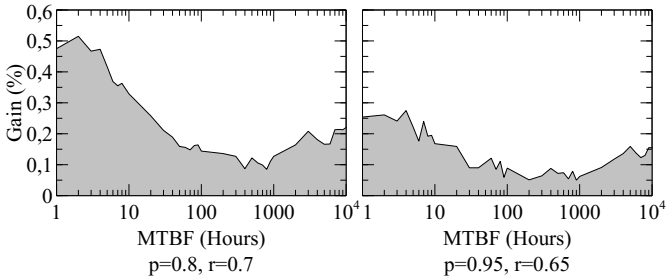


Fig. 4. Time gain percentage of Our Model when compared to Zhu's Model [18] assuming a Variable Checkpointing Overhead ($\alpha = 0.3$, $\beta = 5min$, $R = 10min$)

which is the only model that takes into consideration hybrid fault tolerance among previously compared models. The results are depicted in Fig.3 and Fig.4, and display the gain percentage on the average job finish time when using our model instead of Zhu's model.

5 Conclusion

In this paper, we have brought a new formula for computing the optimum checkpointing interval in systems where hybrid fault tolerance is applied and considering variable size checkpoint dumps. The formula has been compared to state of the art approaches and the results show that our formula brings better results as regards to the job execution time in failure prone environments.

We have considered in this paper that the checkpoint dump size is a function of the execution time and is linear. Therefore, a good direction for future work is considering more general functions for expressing the size of checkpoint dumps.

References

1. Acs - advanced cloud simulator (2014), https://www.github.com/samysadi/acs
2. AReed, D.: High-end computing: The challenge of scale. In: Director's Colloquium, Los Alamos National Laboratory (2004)
3. Bouguerra, M.-S., Trystram, D., Wagner, F.: Complexity analysis of checkpoint scheduling with variable costs. IEEE Transactions on Computers 62(6), 1269–1275 (2013)
4. Daly, J.T.: A higher order estimate of the optimum checkpoint interval for restart dumps. Future Generation Computer Systems 22(3), 303–312 (2006)
5. Egwutuoha, I.P., Levy, D., Selic, B., Chen, S.: A survey of fault tolerance mechanisms and checkpoint/restart implementations for high performance computing systems. The Journal of Supercomputing 65(3), 1302–1326 (2013)
6. Gelenbe, E.: On the optimum checkpoint interval. Journal of the ACM (JACM) 26(2), 259–270 (1979)

7. Gujrati, P., Li, Y., Lan, Z., Thakur, R., White, J.: A meta-learning failure predictor for blue gene/l systems. In: International Conference on Parallel Processing, ICPP 2007, p. 40. IEEE (2007)
8. Li, Y., Lan, Z.: Exploit failure prediction for adaptive fault-tolerance in cluster computing. In: Sixth IEEE International Symposium on Cluster Computing and the Grid, CCGRID 2006, vol. 1, p. 8. IEEE (2006)
9. Ling, Y., Mi, J., Lin, X.: A variational calculus approach to optimal checkpoint placement. IEEE Transactions on Computers 50(7), 699–708 (2001)
10. Ozaki, T., Dohi, T., Okamura, H., Kaio, N.: Distribution-free checkpoint placement algorithms based on min-max principle. IEEE Transactions on Dependable and Secure Computing 3(2), 130–140 (2006)
11. Philp, I.: Software failures and the road to a petaflop machine. In: HPCRI: 1st Workshop on High Performance Computing Reliability Issues. In: Proceedings of the 11th International Symposium on High Performance Computer Architecture, HPCA-11 (2005)
12. Sahoo, R.K., Oliner, A.J., Rish, I., Gupta, M., Moreira, J.E., Ma, S., Vilalta, R., Sivasubramaniam, A.: Critical event prediction for proactive management in large-scale computer clusters. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 426–435. ACM (2003)
13. Salfner, F., Lenk, M., Malek, M.: A survey of online failure prediction methods. ACM Computing Surveys (CSUR) 42(3), 10 (2010)
14. Schroeder, B., Pinheiro, E., Weber, W.-D.: Dram errors in the wild: a large-scale field study. In: ACM SIGMETRICS Performance Evaluation Review, vol. 37, pp. 193–204. ACM (2009)
15. Toueg, S., Babaoglu, Ö.: On the optimum checkpoint selection problem. SIAM Journal on Computing 13(3), 630–649 (1984)
16. Vishwanath, K.V., Nagappan, N.: Characterizing cloud computing hardware reliability. In: Proceedings of the 1st ACM Symposium on Cloud Computing, pp. 193–204. ACM (2010)
17. Young, J.W.: A first order approximation to the optimum checkpoint interval. Communications of the ACM 17(9), 530–531 (1974)
18. Zhu, L., Gu, J., Wang, Y., Zhao, T.: Research on optimum checkpoint interval for hybrid fault tolerance. In: Wu, C., Cohen, A. (eds.) APPT 2013. LNCS, vol. 8299, pp. 367–380. Springer, Heidelberg (2013)

Monitoring Checklist for Ceph Object Storage Infrastructure

Pragya Jain[1], Anita Goel[2(✉)], and S. C. Gupta[3]

[1]Department of Computer Science, University of Delhi, Delhi, India
[2]Department of Computer Science, Dyal Singh College, University of Delhi, India
[3]Department of Computer Science, IIT, Delhi, India
prag_2648@yahoo.co.in,
{goel.anita,gupta.drsc}@gmail.com

Abstract. Object storage cloud is widely used to store unstructured data like photo, emails, video etc. generated from use of digital technologies. The number of object storage services has increased rapidly over the years and so is increased the complexity of the infrastructure behind it. Effective and efficient monitoring is constantly needed to properly operate and manage the complex object storage infrastructure. Ceph is an open source cloud storage platform that provides object storage as a service. Several works have discussed ways to collect the data for monitoring. However, there is little mention of what needs to be monitored. In this paper, we provide an infrastructure monitoring list for Ceph object storage cloud. We analyze the Ceph storage infrastructure and its processes for identifying the proposed lists. The infrastructure monitoring list allows selecting requirements, in contrast to, specifying fresh requirements, for monitoring. The monitoring list helps developer during requirement elicitation of the monitoring functionality when developing a new tool or updating an existing one. The checklist is also useful during monitoring activity for selecting parameters that need to be monitored by the system administrator.

Keywords: Cloud Object Storage · Infrastructure Monitoring · Ceph

1 Introduction

The mass adoption and increasing popularity of digitization technologies has resulted in generation of data in form of videos, photo, blogs, emails, messages, chat data etc. Object storage cloud is a widely adopted paradigm for storing these voluminous data over the Internet. Ceph 0 is open source cloud storage for storing data as object.

The services provided to the subscribers for object storage solutions have rapidly increased and so has complexity of the underlying infrastructure. Monitoring is necessary in cloud to determine health of system and is beneficial for both service provider and consumer 000. There is a need to scale storage nodes, detect and repair failures, manage load surge and improve performance. Due to the elastic nature of cloud, there is a need to constantly monitor the infrastructure at runtime 0 to optimize the use of storage infrastructure with varying demand for storage. Also, disruption in

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 611–623, 2015
DOI: 10.1007/978-3-319-19578-0_50

system performance due to reasons like, node failure, system crash, network error, high memory load etc. requires monitoring during runtime. Moreover, processes for storing activity defined in Ceph require monitoring to detect any erroneous action. Furthermore, since Ceph is integrated with many popular clouds, like, Openstack and Eucalyptus for providing storage as a service, defining of monitoring functionality is essential for determining its proper working.

In Ceph, monitoring functionality is incorporated in different ways – (1) Commands are available to monitor storage cluster, (2) Existing freely available infrastructure monitoring tools, like, Nagios are adapted to suit need of Ceph, or (3) New code is written to include infrastructure monitoring functionality. Generally, freely available infrastructure monitoring tools are used for monitoring Ceph.

Several researchers have discussed different architectures 000000 for monitoring cloud for specific purposes. The focus is mainly on efficient ways of collecting and analyzing data. But, none of them address the issue of what needs to be monitored in object storage cloud. Several tools exist that support monitoring of specific features, like, Calamari monitors Ceph cluster, CollectD and Zabbix monitor system performance, Nagios monitors status of resources, Munin monitors storage capacity. Although the tools specify functionality it supports, there is no mention of requirement specification for the monitoring of Ceph.

In this paper, focus is on creation of the requirement specification for infrastructure monitoring of Ceph from the system administrator perspective. It helps during development of tools for the system administrator, in choosing and specifying requirements for the monitoring functionality.

Here, a infrastructure monitoring checklist is presented that facilitates in selecting requirement when developling tools and techniques for monitoring of Ceph. We have classifed the infrastructure monitoring into four components, namely, (1) Background process functionality, (2) Storage infrastructure attributes, (3) Storage usage data, and (4) OS process utilization data. The monitoring checklist is defined for the four identified components of Ceph. The checklist is for both the administrator and the developer, and facilitates during requirement elicitation in identifying the monitoring functionality to be included in a tool. During requirement phase of tool development, functionality needed for monitoring of Ceph can be selected from the checklist.

For understanding requirements of monitoring Ceph and for formulating the checklist, a study of architecture of Ceph, processes in Ceph for storing and managing data, monitoring commands and configurable parameters of Ceph was performed. A study of associated plug-in of some standard open source monitoring software was also performed. This collectively defines understanding storage architecture and available monitoring provisions for Ceph. Using the use-case based approach; the requirements for monitoring have been identified from the system administrator perspective. The components of infrastructure monitoring, based on interaction of system administrator with Ceph infrastructure are defined. The functionality of each identified component of infrastructure monitoring has been identified.

The monitoring checklist allows selecting requirements from the checklist, in contrast to, specifying fresh requirements, when developing new tool for monitoring.

From the checklist, all or part of functionality may be selected. The checklist is for use during requirement elicitation phase, and also for validation and verification of requirements during testing phase of the tool development. The requirement checklist presented here can be easily updated to include any new functionality or feature.

The Ceph infrastructure checklist presented here has been applied to three popular infrastructure monitoring tools of Ceph - Calamari 0, Nagios [18], and CollectD 0 to identify monitoring functionality provided by them. The work is being currently extended to provide generic infrastructure monitoring list for cloud object storage.

In this paper, section 2 gives an overview of Ceph object storage. Section 3 describes Ceph monitoring commands and configurable parameters. Section 4 discusses the components of infrastructure monitoring. Section 5 describes the monitoring checklist in detail. Section 6 illustrates few examples on which the checklist has been applied. Section 7 lists benefits of using the infrastructure monitoring list. Section 8 is a survey of related work. Section 9 states the conclusion.

2 Ceph Object Storage

The Ceph 0 object storage architecture comprises of three main components – Radosgw, Librados and RADOS.

Radosgw is a client interface for object storage that allows end-user to store and retrieve data. It supports Swift and S3 compatible APIs for facilitating end-user to perform various operations, such as, create, read, update and delete data as an object.

Librados is storage cluster protocol that provides native interface to interact with storage cluster and supports different languages, like, C, C++, Java, Ruby and Python. It allows client to interact with Ceph storage cluster, directly, using the defined API.

RADOS 0 is a reliable, autonomous and distributed object store. It consists of two sub-components – Monitor and OSD (Object Storage Device) Daemon. Monitor maintains current status of each component of cluster. Usually, one monitor is sufficient for this purpose, but to ensure high availability, a few monitors are used and a quorum for consensus about current state is established among them. OSD daemon is responsible for reading/writing data to/from storage cluster. OSD daemons communicate with each other to check whether other OSDs are in up and running state and also to replicate data.

3 Ceph Monitoring Commands and Configurable Parameters

In Ceph, several commands 0 exist that provide health of storage cluster and state of individual components, like, their running status and condition. There are also some commands that provide usage statistics of storage cluster.

The component of Ceph storage cluster has some configurable parameters that can be set according to the need of cluster. These parameters can be set at the time of software installation or can be changed dynamically at runtime. Ceph stores its configurable parameters in its configuration file. The configurable parameters 0 are divided into four major sections – global, osd, mon and client.radosgw. Configurable

parameters set in 'global' section are applied to all instances of all components. Parameters set in 'osd', 'mon' and 'client.radosgw' is applicable for instances belonging to OSD daemon, monitor and Radosgw, respectively. Configurable parameters define working of different processes running for the component.

4 Components of Infrastructure Monitoring

The classification of Ceph infrastructure into different components provides a framework for categorization of functionality of monitoring, from the administrator perspective. To understand requirements of infrastructure monitoring, a study of the components and processes executing on them has been performed. An in-depth study of different monitoring commands and configurable parameters of Ceph object storage has also been done. The infrastructure monitoring has been divided into four broad components as follows:

- *Background Process* - functionality of processes running in background
- *Storage Infrastructure* - attributes of storage infrastructure
- *Storage Usage* - utilization of storage infrastructure
- *OS Process Utilization* - utilization of OS processes

For Ceph, the authors define infrastructure monitoring as, "Monitoring physical infrastructure, logical infrastructure and associated processes". The components of infrastructure monitoring are briefly described in the following subsections.

4.1 Background Process

During working of the Ceph object storage, several processes run in the background to perform the tasks defined in Ceph. From the different processes present in Ceph software, we identified the processes that are required to be monitored during runtime as shown in Fig. 1. These processes are required to be monitored to check health of system. The background processes that are required to be monitored are - Heartbeat, Authentication, Data scrubbing, Peering, Backfilling, Recovery and Synchronization.

Heartbeat ensures that OSDs responsible for maintaining copies of data are in up and running states. OSDs check heartbeat of other OSDs periodically and report the status to monitor.

Authentication is used to authenticate and authorize the client accessing the storage. Monitor is responsible for authentication process. Client can have different rights for access, like, read-only, write, access to admin commands, etc.

Data scrubbing checks data integrity. The process runs on OSDs and compares objects with their replica stored in another OSD. There are two types of scrubbing– light scrubbing and deep scrubbing. In light scrubbing, metadata of objects is compared to catch bugs. In deep scrubbing, data in objects is compared bit-by-bit. Usually, light and deep scrubbing are performed daily and weekly respectively.

Peering is required for creating an agreement about state of all objects among OSDs that are responsible to keep copy of objects before replication.

Synchronization ensures availability of data in a federated system implemented with multiple regions and multiple zones. A cluster must have a master region and a region must have a master zone. Synchronization process runs on Radosgw. There are two types of synchronization - data synchronization and metadata synchronization. In data synchronization, data of master zone in a region is replicated to a secondary zone of that region. In metadata synchronization, metadata of users and buckets is replicated from master zone in master region to master zone in a secondary region.

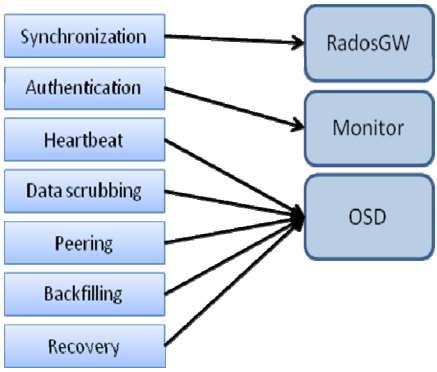


Fig. 1. Background processes for Infrastructure monitoring in Ceph

Backfilling runs when OSD is added or removed to/from Ceph storage cluster. In order to rebalance cluster, objects are moved to or from OSDs. This migration takes place as 'backfilling' at lower priority to maintain operational performance of system.

Recovery runs when OSD crashes and comes back online. In such condition, several objects stored in OSD get outdated and goes in recovery mode, when it restarts. To maintain operational performance of system, recovery process takes place with some limitations.

Several other processes like logging and journaling do not require monitoring during runtime.

4.2 Storage Infrastructure

The storage infrastructure of Ceph is logically divided into clusters which contain few monitors and a large number of OSDs. In a typical scenario, an OSD maps to a storage drive or a RAID group. The storage cluster is divided into pools, which are further divided into Placement Group (PG). Each PG maps to some OSDs.

A pool facilitates segregation of data, logically, based on user's requirement. For providing availability, pool is specified as replicated or erasure-coded. Replicated pool maintains multiple copies of data. In erasure-coded pool, data is divided into number of chunks associated with some code chunks. The data is stored in PGs within a pool. For fault tolerance, each copy of PG is stored in separate OSD. A set of OSDs that are responsible for keeping copy of a PG is called Acting set of that PG and a set of OSDs that are ready to handle incoming client request is Up set of that PG.

Generally, Acting set and Up set of a PG are identical. If they are not found identical, it implies that Ceph is migrating data or an OSD is recovering or there is any problem. One OSD in Acting set is primary OSD. Client communicates with primary OSD to read/write data. Primary OSD interacts with other OSDs to replicate data.

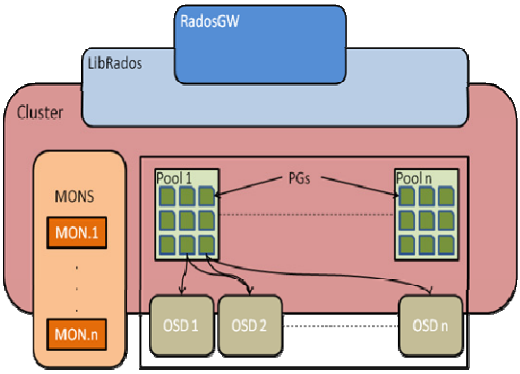


Fig. 2. Ceph Object Storage Structure

Fig. 2 shows object storage infrastructure for Ceph. Monitoring is required at all the different levels of physical and logical infrastructure, for observing existing resources. It helps when there is a need to add or remove resources and to detect failing or failed drives for replacement.

4.3 Storage Usage

The amount of the logical and physical infrastructure that is being consumed is required to be monitored to identify full or near full storage. The logical infrastructure is monitored at different levels - cluster, pool and PG level. This helps in scaling up and scaling down the system resources.

4.4 OS Processes Utilization

The Ceph object storage uses CPU, memory, and network for its own working. Different processes, like, heartbeat, peering etc. run on different components of storage cluster and utilize CPU, memory and network. OS processes utilization is required to be monitored to improve efficiency and performance of system.

5 Infrastructure Monitoring List

For arriving at the monitoring checklist, the Ceph infrastructure was classified under four components - Background process functionality, Storage infrastructure attributes, Storage usage data and OS Process utilization data.

Background process checklist consists of functionality running in background of Ceph. During requirement elicitation, this list helps in deciding process parameters that need to be monitored for Ceph software. The background processes are monitored for three entities – monitor, OSD, and Radosgw.

Heartbeat and peering processes are required to be monitored for finding OSDs in 'Up' and 'Acting' set of a PG, respectively, to check if number of OSDs in Acting set are same as that defined in pool size and OSDs in Up set of a PG are equivalent to OSDs in Acting set of that PG. The users are monitored for authentication to check access permissions according to defined capabilities. Data scrubbing needs to be monitored for type of scrubbing, its frequency, and number of pending scrubs and errors to identify rate of corrupted files found in system so that any abnormality can be identified. Data and metadata synchronization have parameters, such as, errors during sync, wait time, count of shards that are checked or failed, and error listing metadata to determine correct working of system. Backfilling and recovering processes are monitored for their respective status so that impact on system performance can be decreased. Table 1 lists requirement checklist for background process functionality.

Storage Infrastructure checklist defines parameters for logical and physical infrastrcuture that needs to be monitored. During requirement elicitation this list helps in deciding parameters of infrastructure that need to be monitored for Ceph. The storage infrastructure is divided into five levels – cluster, monitor, OSD, pool and PG.

Table 1. Background Process Functionality

Process	Parameters	Monitor	OSD	Radosgw
Heartbeat	OSDs in Up set of a PG	-	✓	-
Authentication	Users with different capabilities	✓	-	-
Data scrubbing	Type – Light/Deep, frequency, scrub pending, no. of errors	-	✓	-
Peering	OSDs in Acting set of PG	-	✓	-
Data synchronization	Sync error, incremental sync error, retry wait time/until next sync, object sync timeout, no. of shards to check/failed, no. of items synced successfully/ processed	-	-	✓
Metadata synchronization	Time to wait for bucket log consistency, Error listing metadata	-	-	✓
Backfilling	Count, frequency, time to wait for retrying	-	✓	-
Recovering	No. of active recovery request/ recovered chunks, time to delay	-	✓	-

At cluster level, parameters are identified as cluster health status, number and list of monitors, OSDs, pools and PGs in cluster. Detail of each OSD, current epoch, and OSD status can be monitored at OSD level. Pool level parameters are number of PGs in pool, pool is replicated or erasure coded etc. At PG level, parameters define state of PG. Table 2 lists requirement checklist for storage infrastructure attributes.

Storage usage checklist consists of parameters that provide data about the usage of storage infrastructure. During requirement elicitation the list helps in deciding parameters for usage of storage infrastructure that need to be monitored. Storage usage data for monitoring is defined at three levels - cluster, pool and PG.

IOPS (Input Output Per Second) measure input/output load to avoid I/O bottleneck in system. Latency provides time taken in data transfer so that in cases of interruption the cause can be found.

Table 2. Storage Infrastructure Attributes

Level	Parameters
Cluster	Cluster ID
	Cluster health status
	Number and list of monitors
	Number and list of OSDs
	Number and list of Pools
	Number and list of PGs
Monitor	Detail – position, name, address, port of monitor
	Current epoch – when map created, last modified
	Status - Running/ not running
	Status of monitor quorum
OSD	Details – id, weight, type, name
	Current epoch – when map created, last modified
	Status - In/out, up/down
Pool	Details - Name, Pool ID
	Number and list of PGs
	Replicated/erasure coded
	Cache tiering status
Placement	Detail – PG ID, PG version, timestamp
Group	PG state (Creating, Peering, Active, Clean, Degraded, Recovering, Backfilling,
	Remapped, Stale, Unclean, Inactive)

Total storage capacity and free space availble are inspected so that alerts can be raised before system reaches near-full capacity. Amount of data stored and number of objects stored provide estimate of storage capacity. IOPS and latency are monitored at cluster and pool level. Notional value monitored at pool level determines utilized space excluding space used by its replicas. Table 3 lists storage infrastructure usage parameters at different levels.

OS processes checklist contains parameters to determine utilization of operating system processes. During requirement elicitation this list helps in deciding parameters for utilization of OS processes that need to be monitored.

CPU, memory and network utilization data is monitored to determine consumption of OS resources during execution. It helps to identify processes and components that are under utilizing or highly utilizing OS resources so that extra resources can be provisioned based on demand. Table 4 lists parameters for OS process utilization.

Table 3. Storage Usage Data

Level/Parameters	Cluster	Pool	PG
IOPS – read, write	✓	✓	-
Latency – max., avg., min.	✓	✓	-
Overall storage capacity	✓	-	✓
Amount of data stored	Total & Notional	Notional	Total
Number of objects stored	Total	Notional	-
Amount of free space available/ used	Total	Notional	Total

Table 4. OS Process Utilization data

Parameters	Reason
CPU Utilization	Find CPU consumption by processes and system to identify processes and components which have high CPU load and which are under utilized
Memory utilization	Track available memory to determine processes and components that are consuming more memory so that memory can be upgraded
Network utilization	Track network traffic and identify network interfaces that have excessive use

6 Case Study

The infrastructure monitoring functionality lists have been applied for case study to three monitoring software - Calamari 0, Nagios 0, and CollectD 0.

Calamari is management and monitoring service specifically for Ceph. It exposes high level REST APIs and a user interface built on these APIs for monitoring Ceph infrastructure. *Nagios* is open source infrastructure monitoring software that enables organizations to identify and resolve IT infrastructure problems before they have drastic effect on system. Nagios provides some built in plug-ins for monitoring health of cluster and individual components of Ceph object storage, like, check_ceph_health and check_ceph_mon. *CollectD* is daemon that collects system information and helps system administrators to maintain an overview of resources to avoid bottlenecks.

Table 5 displays comparative checklist of the three monitoring software for storage infrastructure, usage and OS process utilization of infrastructure monitoring. In the table, '√' denotes parameter is supported by tool; 'x' not supported. 1st, 2nd and 3rd column for each is for storage infrastructure, usage and OS process, respectively.

Fig. 3 displays percentage of parameters of each monitoring checklist functionality supported by tools in our case study. Some of our key observations are as follows-

- *Calamari* monitors 10% of background processes; 95.65 % of storage infrastructure; 87.5% storage usage; and 33.33% OS processes.
- *Nagios* monitors mainly the status of storage infrastructure (69.56%). It does not monitor background processes, storage usage and OS processes.

- *CollectD* monitors 100% storage usage; 100% OS processes utilization; 38.46% of storage infrastructure. It does not monitor background processes.

Some interesting observations emerging from the case study are as follows-

- Background processes is only monitored 10% by Calamari
- Storage infrastructure is monitored by all three - Calamari, Nagios, CollectD
- Storage usage is monitored by Calamari and CollectD
- OS processes is monitored by Calamari and CollectD

Table 5. Storage infrastructure, Storage usage, and OS processes case study

Level	Parameters–Storage Infrastructure	Parameters–Storage Usage	Parameters–OS Processes	Calamari			Nagios			CollectD	
Cluster	ID	IOPS – read, write	CPU Util.	√	√	√	√	x	x	x	√
	Health status	Latency-max/avg/min	Memory Util.	√	x	x	√	x	x	x	√
	No. & list of monitors	Notional data stored	Network Util.	√	√	x	√	x	x	√	√
	No. & list of OSD	No. of objects stored	-	√	√	-	√	x	-	√	√
	No. & list of Pools	Total data stored	-	√	√	-	√	x	-	√	√
	No. &List of PGs	Free space available	-	√	√	-	√	x	-	√	√
	-	Used raw storage	-	-	√	-	-	x	-	-	√
	-	% of raw storage used	-	-	√	-	-	x	-	-	√
	-	Overall storage capacity	-	-	√	-	-	√	-	-	√
Monitor	Detail	-	-	√	-	-	√	-	-	x	-
	Current epoch	-	-	√	-	-	√	-	-	√	-
	Status	-	-	√	-	-	√	-	-	√	-
	Monitor quorum status	-	-	√	-	-	√	-	-	√	-
OSD	Detail	-	-	√	-	-	√	-	-	x	-
	Current epoch	-	-	√	-	-	√	-	-	x	-
	Status- in/out, up/down	-	-	√	-	-	√	-	-	√	-
Pool	Name, ID	IOPS – read, write	CPU Util.	√	√	√	x	x	x	x	√
	No. & list of PGs	Latency – max/ avg/min	Memory Util.	√	x	x	√	x	x	√	√
	Replicate/Erasure	Notional data stored	Network Util.	√	√	x	x	x	x	x	√
	Cache tiering status	Notional objects stored	-	√	√	-	x	x	-	x	√
PG	Detail	Amount of data used	-	√	√	-	x	x	-	x	√
	-	Free storage capacity	-	-	√	-	-	x	-	-	√
	PG state	Total storage capacity	-	√	√	-	√	x	-	√	√

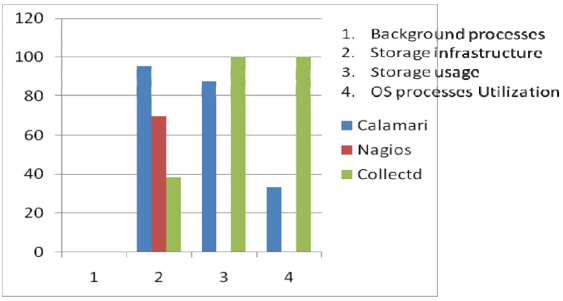


Fig. 3. Graph showing percentage of checklist used by monitoring software

It can be seen that tools offer different coverage for components being monitored and there is no consistency for same. Also background processes are hardly monitored because Ceph does not have commands to provide status of running processes.

7 Benefits of Checklist

The requirement checklist presented here has been derived after performing a detailed study of Ceph storage architecture, processes running in it and exhaustive study of basic monitoring commands and configurable parameters over Ceph.

Usually, infrastructure monitoring software is developed for a specific purpose without planning or preparation of list of possible functionality that can be included. The checklist helps system administrator to choose functionality required to monitor infrastructure of Ceph with minimum effort. The developers use checklist to check functionality required for infrastructure monitoring during development of tool. The checklist allows developer and administrator to include more functionality in monitoring software rather than just basic functionality.

8 Related Work

Cloud monitoring has gathered focus in research. Several researchers discuss about research motivation, approaches used for monitoring cloud and different methodologies applied to monitor a cloud for different purposes. Alhamazani et al. 0 discusses methodology to monitor cloud for facilitating automated QoS management; Adinarayan 0 discuss challenges in monitoring private cloud and describe capabilities of IBM SmartCloud monitoring to tackle these challenges.

Several frameworks are proposed by researchers for different purposes on monitoring the cloud infrastructure 000. Gogouvitis et al. 0 propose an architectural design and implementation of monitoring solution in context of VISION cloud project. Mdhaffar et al. 0 propose dynamic Complex Event Processing architecture for cloud monitoring and analysis; Uriate and Westphall 0 propose monitoring architecture ‘Panoptes’ for autonomic clouds; Chaves, et al. 0 discuss design and implementation of private cloud monitoring system (PCMONS).

The frameworks and architectures highlight ways of collecting data from system required for monitoring and how to monitor. However, there is no mention of parameters required to be monitored in cloud. Usually, freely available monitoring software, like, Zenoss, Nagios are adapted for incorporating monitoring functionality for Ceph object storage. Our extensive search for work carried out for finding list of parameters required for monitoring Ceph object storage yielded no result.

9 Conclusion and Future Work

In this paper, we have presented infrastructure monitoring list for Ceph object storage. The list eases the task of administrator and developers by providing them a list from where the functionality can be selected. Designers and developers of new monitoring software for Ceph can also use the list as a reference for identifying possible functionality that can be incorporated in monitoring software. The list is extendible and can be updated to add new functionality and features.

Since our functionality checklist is specific for Ceph object storage, other cloud object storage may have some more functionality which does not lie in scope of this paper. In future, the authors aim to develop a generic functionality checklist for cloud object storage system. We also propose to prioritize the proposed list.

References

1. Adinarayan, G.: Monitoring and Capacity Planning of Private Clouds: The Challenges and the Solutions. In: IEEE Int. Conf. on Cloud Computing in Emerging Markets (CCEM), India, pp. 1–3 (2012)
2. Alhamazani, K., et al.: Cloud monitoring for optimizing the QoS of hosted applications. In: IEEE 4th Int. Conf. on Cloud Computing Technology and Science (CloudCom), Taipei, pp. 765–770 (2012)
3. Barbosa de Carvahlo, M., et al.: A cloud monitoring framework for self-configured monitoring slices based on multiple tools. In: 9th Int. Conf. on Network and Service Management (CNSM), Zurich, pp. 180–184 (2013)
4. Chaves, S., et al.: Towards an architecture for monitoring private clouds. IEEE Communications Magazine 49, 130–137 (2011)
5. Gogouvitis, S., et al.: A Monitoring Mechanism for Storage Clouds. In: 2nd Int. Conf. on Cloud and Green Computing, Xiangtan, pp. 153–159 (2012)
6. Grobauer, B., Walloschek, T., Stocker, E.: Understanding cloud-computing vulnerabilities. IEEE Security and Privacy 9, 50–57 (2010)
7. Mdhaffar, A., et al.: A Dynamic Complex Event Processing Architecture for Cloud Monitoring and Analysis. In: IEEE 5th Int. Conf. on Cloud Computing Technology and Science (CloudCom), Bristol, vol. 2, pp. 270–275 (2013)
8. Moses, J., Iyer, R., Illikkal, R., Srinivasan, S., Aisopos, K.: Shared Resource Monitoring and Throughput Optimization in Cloud-Computing Datacenters. In: IEEE Int. Parallel & Distributed Processing Symposium (IPDPS), Anchorage AK, pp. 1024–1033 (2011)
9. Rehman, Z., et al.: A Framework for User Feedback based Cloud Service Monitoring. In: 6th Int. Conf. on Complex, Intelligent and Software Intensive Systems (CISIS), Palermo, pp. 257–262 (2012)

10. Shao, J., Wei, H., Wang, Q., Mei, H.: A Runtime Model Based Monitoring Approach for Cloud. In: IEEE 3rd Int. Conf. on Cloud Computing (CLOUD), Miami, FL, pp. 313–320 (2010)
11. Uriarte, R., Westphall, C.: Panoptes A monitoring architecture and framework for supporting autonomic Clouds. In: IEEE Network Operations and Management Symposium (NOMS), Krakow, pp. 5–9 (2014)
12. Weil, S., et al.: RADOS A Scalable, Reliable Storage Service for Petabyte-scale Storage Clusters. In: 2nd Int. Workshop on Petascale Data Storage, pp. 35–44. ACM, New York (2007)
13. Yongdnog, H., et al.: A Scalable And Integrated Cloud Monitoring Framework Based On Distributed Storage. In: 10th Web Information System and Application Conference, Yangzhou, pp. 318–323 (2013)
14. https://github.com/ceph/calamari
15. http://ceph.com/ceph-storage/object-storage/
16. https://collectd.org/
17. http://ceph.com/docs/v0.78/rados/operations/monitoring/
18. http://www.nagios.org/

Towards a Formalization of Real-Time Patterns-Based Designs

Kamel Boukhelfa(✉) and Faiza Belala

Department of Software Technologies and Information Systems, Faculty of New Information Technologies and Communication, University of Constantine 2, Ali Mendjeli, Algeria {kamel.boukhelfa,faiza.belala}@univ-constantine2.dz
http://www.univ-constantine2.dz

Abstract. Informal description (UML and text) of design patterns is adopted to facilitate their understanding by software developers. However, these descriptions lead to ambiguities, mainly when we consider Real time Design Patterns that deal with critical problems encountered in the design of real-time systems. Hence, there is a need for formal specification of the DPs and RTDPs to insure their successful application. In this paper, we propose a formalization approach of the system design based on real-time patterns (RTDPs). The processes of instantiation and composition of design patterns, permit us to generate design models (structural and dynamic) of complex systems. The resulting designs are represented in UML-MARTE profile to express the temporal properties and constraints. The algebraic specifications (in Maude language) become more natural and more efficient.

1 Introduction

A design pattern expresses solution of a known and recurrent problem in a particular context [5]. Design patterns are applied in object programming software to improve the quality of the resulting system. The reuse concept is also important in the development of real-time and embedded systems. Thus, design patterns can be used to capture the experience and allow the reuse of the "good" solutions to resolve the problems encountered during the design process of such systems [3]. Intuitively, the term "real-time" refers to design patterns those dealing with the temporal aspects of systems, whereas this is not always the case. Indeed, real-time design patterns deal with the general problems encountered in the design of real-time systems (that may be or not related to the time) such as synchronization or memory allocation. The real-time design patterns vary according to their areas of application and according to the design approaches. Generally, design patterns and also real-time design patterns were described, until now, by using a combination of textual descriptions, object oriented graphical notations such as UML diagrams and sample fragments of code [5], [3]. This informal description of design patterns is adopted to facilitate their understanding by software developers. However, formal specifications provide a precise and rigorous description for better understanding patterns and their instantiation and composition. This

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 624–635, 2015.
DOI: 10.1007/978-3-319-19578-0_51

description is then ready for several analysis and verifications upon one or more functional or non-functional properties.

Several research work around design patterns deal with issues related to their representation and specification. We distinguish two points of view adopted for this purpose. The first one concerns all works that adopt the meta-modeling approaches and consequently the definition of patterns modeling languages based on UML. These works aim in general to provide solutions for integrating design patterns in CASE tools. The second kind of research work is characterized by the use of the formal methods to specify the design patterns and then provide suitable models to the analysis and verification stages. However, few studies are particularly interested in RTDPs. In this work, we start from the real-time design patterns as the basic models. Through the instantiation and the composition processes, we conceive design models and represent them in UML-MARTE profile [10]. We use Rewriting Logic [9] as a formal foundation for the specification of the Pattern-Based models and thus, we encode in Maude language [2] the formal specification of both parts of those models, namely the structural and dynamic part.

The rest of this paper is organized as follows: After recalling the used basic concepts of RT Design Patterns, MARTE profile and rewriting logic via its practical language Maude in section 2, we outline in section 3, how it is possible to give a formal base to real-time systems designs thanks to a judicious coupling of UML-MARTE profile and rewriting theories. Then, in section 4, we describe the formalization approach with Maude's object-oriented modules, through a realistic example. Finally, we conclude the paper with constructive remarks and future work.

2 Basic Concepts

2.1 Real-Time Design Patterns

In object oriented programming, design patterns are considered as a mean to encapsulate the knowledge of experienced software designers and represented it in an understandable form in order to permit its reuse. For each design pattern, are defined the roles of classes, relationships between classes and objects, and how this pattern can be applied to resolve a given problem in a specific context. The structure describing a design pattern mainly includes the name, problem, solution and consequence [3]. RT design patterns are a kind of patterns that have evolved specifically for real-time systems, and they provide various approaches to addressing the fundamental real-time scheduling, communications, and synchronization problems [3]. As a GOF design patterns, RTDPs are represented in UML and the most temporal constraints (especially in the interactions) are expressed in the natural language.

2.2 The UML Profile for MARTE

The UML profile for MARTE (Modeling and Analysis of Real-Time and Embedded systems) is an OMG standard. It provides support for specification, design and verification/validation stages. This new profile is intended to replace the existing UML Profile for Schedulability, Performance and Time [10]. Model-based design of RTE systems with MARTE proceeds mostly in a declarative way. The users can annotate their models with real-time or embedded concerns using the extensions defined within the HLAM (High-Level Application Modeling) sub-profile (see the next section). The HLAM package provides possibilities of modeling on one hand quantitative features such as deadline, period and, in the other hand, qualitative features that are related to behavior, communication and concurrency. MARTE provide the NFP package (Non-functional Properties Modeling) in order to specify the NFP of properties in a detailed way [10].

2.3 Rewriting Logic and Maude

Rewriting logic (RL) is known as being the logic of concurrent change, taking into account the state and the concurrent systems calculus. It is shown as a unifying semantic framework of several concurrent systems and models [9]. In RL, a dynamic system is represented by a rewriting theory $\mathcal{R} = (\Sigma, E, R, L)$, describing the complex structure of its states and the various possible transitions between them. The theoretical concepts of the rewriting logic are implemented through the Maude language [9,2] that integrates object oriented programming, used in our formalization to encode the DPs and their meta-models specifications. Maude logical basis gives a clear definition of the object oriented semantics and makes it a good choice for the formal specification of object oriented systems.

3 Formalization Approach Principle

First, we use a given design pattern to generate an UML design (structural and dynamic parts). The resulting design will be enriched by the MARTE notations, namely the concepts defined in HLAM sub-profile, such as `RtUnit`, `PpUnit` and `Rtfeature`, and those defined in the NFP sub-profile, such as `NFP_DateTime`, `NFP_Duration` and `NFP_Frequency`. The second step allows to transcript UML-Marte description to Maude specification. Here, we use Full-Maude, an extension of Maude, that allows us to manipulate the object-oriented concepts, especially objects, classes and attributes. We show in the following sub-sections, how we encode, any system design, described with UML-MARTE and RT Design Patterns coupling, in Maude.

3.1 Static Part

For the structural part, we can note the existence of a correspondence between some concepts of Maude language and UML-MARTE concepts. Unfortunately,

this correspondence is not fully established, there are various concepts in UML-MARTE with no direct equivalent in Maude. The structural part of a design pattern is represented as an UML classes diagram and serves as a model to generate, by means of the instantiation mechanism any structural design based on this pattern. The table 1 contains the MARTE concepts and their correspondences in Maude. For some MARTE concepts without direct correspondence, we also propose their definitions in Maude. For the stereotyping, we define a new class for each stereotype and so, the stereotyped class (in MARTE) is represented by a subclass in Maude. While, for the specification of the methods definition within classes, we define a new `sort` called `Method` and we add the declaration of a Maude operation that permits to link each method to its appropriate class (`op Methods : class -> SetMethod)`. In addition, we use the predefined concepts in several modules of Maude such as the `SET` module, for defining empty and non-empty set (`Set`, `NeSet`), and others modules such as `BOOL`, `FLOAT`, `NAT` and `STRING` to express respectively the types Boolean, Float, Natural and string of characters.

Table 1. Correspondence between MARTE and Maude concepts

MARTE Concept	Maude Concept
Class/objet	Class/Oid
Attribute	Attribute
Directed Association	Operation
Non-Directed Association	Two operation (one for each direction)
Association 1..1/1..*/1..n	operation /op − > Set / op − > NeSet (not empty Set)
Composition	Operation
inheritance	Subclass

3.2 Dynamic Part

The dynamic part of a design pattern represents the interaction between different objects instantiated from classes that form a pattern-based design. This part is often represented by a sequence diagram with all the interactions between objects, shown as signals. Firstly, we declare a new sort called `Signal` that expresses the interaction between two objects. Secondly, we define an operation `Instance` that represents the objet creation signal. Thus, we can specify all objects related to a given activity execution (represented as sequence diagram). The objects can be declared at the start of this activity (of `Oid` type) or created during the execution.

3.3 Real-Time Features

MARTE provides Real-time unit concept (*RtUnit*) defined in HLAM package. An *RtUnit* may be seen as an autonomous execution resource, able to handle different messages at the same time. It can manage concurrency and real-time

constraints attached to incoming messages [10]. Any real-time unit can invoke services of other real-time units, send signals or data without worrying about concurrency issues. Another important point to consider when modelling concurrency system is to be able to represent shared information. For that purpose, MARTE introduce the concept of protected passive unit (*PpUnit*). PpUnit specify the concurrency policy units either globally for all of their provided services (`concPolicy` attribute), or locally through the `concPolicy` attribute of an `RtService`. We will stereotype the classes as `RtUnit` or `PpUnit` regarding their role in the design model. However, operations can be stereotyped as `RtService` for example. We can add the `rtf` stereotype at the methods dealing with real-time features such as deadline and reference time. The temporal constraints are expressed in OCL (Object Constrained Language) for instance, a maximum time to perform an activity. For the occurrence kind of a signal (`occkind`), we define a Maude operation called `periodic` that permits to identify the nature of this signal appearance (periodically or not). In the case of a periodic signal the `periodVal` operation is defined to get the value of the period. For the simplicity, we consider the default unit of time (ms). The others elements characterizing a signal are represented in Maude language as operations upon this signal. The temporal constraints represent the conditions on the actions that need to be satisfied, they are expressed in the OCL language (Object Constraints Language). In addition, we define two sorts, `Time` and `Value` to specify the temporal variables (eg. triggering instants of signals) and their values. Consequently, it is necessary to have an operation to get the value of an instant t (`Rvalue`) and a conditional equation to check whether the imposed constraints is verified or not (`Satisfy`).

4 Running Example: A "Cruise Control System"

This system controls and regulates the speed of a car according to the encountered situations (obstacle, car ahead too closely, etc.). The controller requires the services of three types of sensors, a Speed Sensor, a Laser device to calculate the distance between the car and obstacles and a radar to detect possible obstacles. For simplicity, only the Speed Sensor is considered.

4.1 System Modelling in MARTE

For modelling the system, we use `Observer` and `Sensor` patterns and we compose their instances to generate the structural design of the system. The composition is achieved in a simple way, namely through the overlapping of common elements in the two instances.

The problem addressed by the `Observer` Pattern is how to notify some number of clients in a timely fashion of a data value according to some abstract policy, such as "when it changes," "every so often," "at most every so often," and "at least every so often" [3]. The basic solution offered by the `Observer` pattern is to have the clients "subscribe" to the server to be notified about the value in question according the defined policy.

A Speed Sensor is defined as a device that measures or detects a physical phenomenon (temperature, pressure, speed, etc.) and transmits the measure values at real-time to the command ends. The RT-Design pattern `Sensor` [1] can be specialized as possible types of sensors : `Active Sensor`, `Passive Sensor`, `Fixed Sensor` and `Mobile Sensor`. We use `Active Sensor` pattern which is able to send signals `Setvalue` to one or more objects for modifying the measured value. The class `measure` stores the data taken by the Sensor, while the attributes (`timestamp`, `validity duration`) are used to represent the characteristics of real-time data supported. The class `Observed element` is used for the physical supervised device description (a wheel for example).

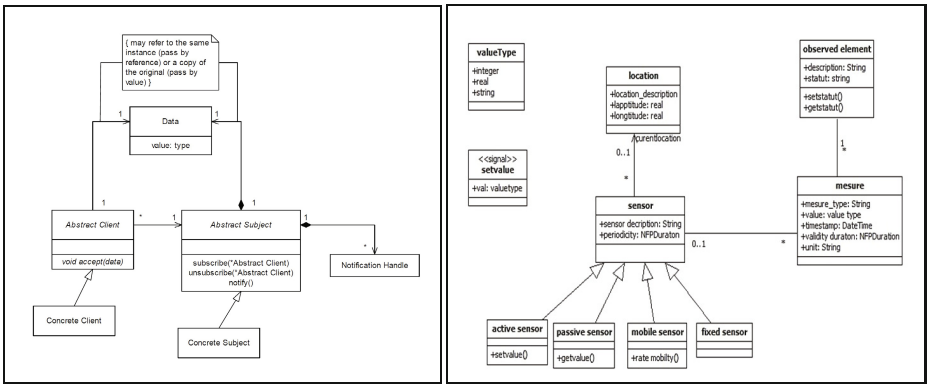


Fig. 1. "Observer" and "Sensor" Real-Time Design patterns Structures

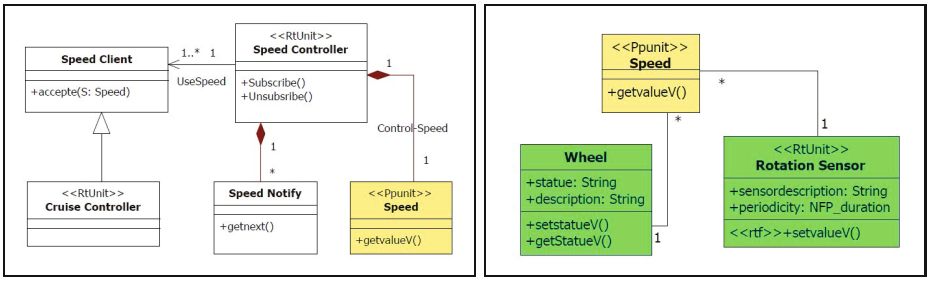


Fig. 2. Instances of the "Observer" and "Sensor"

For the structural design modelling, we use two instances of the pattern `Observer` to model the structure of the sub-systems (*Speed Controller* and *Distance Controller*). For each sub-system, we need to instantiate `Observer` and `Sensor` patterns and so, compose these instances.

We use an instance of `Observer` pattern to model the *Speed Controller* subsystem. The resulting model is represented in MARTE and enriched with temporal and NFP proprieties. In the same way, we proceed for modelling *Distance Controller* sub-system (Fig. 2). Similarly, we use two instances of the `Sensor` RT-Design pattern (Fig. 2) to model the capturing of the speed rotation of a car wheel, the detection of the possible obstacles in front of a car and the distance measure which separate them from the car (Laser device). The composition of the instances of `Observer` and `Sensor` patterns respectively regarding the common elements (the `Speed` class in first case, and `Distance` class in the second one) produces the design model of the complete system. In the dynamic

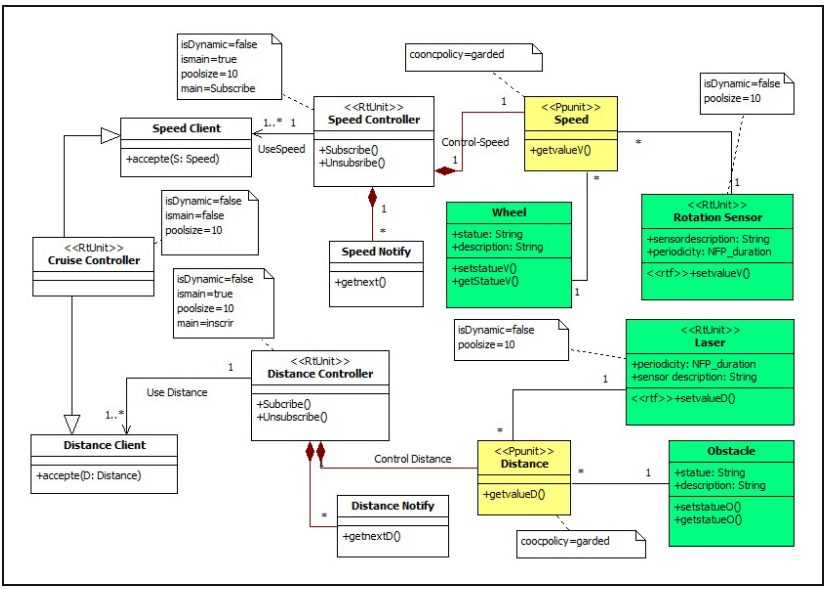


Fig. 3. Structural part of the *"Cruise control system"* in MARTE profile

design modelling, we describe the system by using a sequence diagram. This diagram shows a scenario of data acquisition and how the system will react to synchronous or asynchronous events. The interactions that have temporal properties are stereotyped as `RtFeature`. This allows us to model the temporal behavior of these interactions (occurrence mode, deadline, etc.). MARTE profile also allows us to set time restrictions upon interactions with "time constraint" (eg. `t2-t1<(5ms))`. The figure 4 shows the sequence diagram for the *Cruise Control System* to perform the task of capturing the car speed and the distance in the case of a nearly car. The cruise control object needs two services (internal speed and distance), so it must subscribe into both lists of notification.

At the time t1[i] for example (each action i starts at t1[i] to get a speed value), a message is sent by the speed controller object. This message represents a call

to `getvalue` method. The message is stereotyped by `RtFeature` to represent temporal properties such as the period of the occurrence of this message (20 ms). It will be followed by other interactions for notifying the clients. These interactions must be completed at time t2 with a maximum delay of 5 ms.

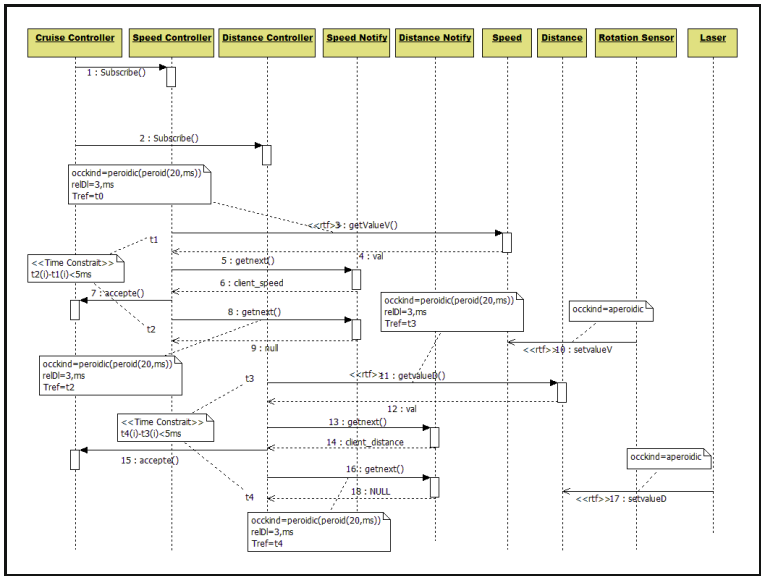


Fig. 4. Dynamic part of the "*Cruise control system*": Capturing a speed and distance Activity

4.2 Maude Formalization

The major advantage of the rewriting logic and its language Maude is its ability to specify in the same formalism both, the structural and the dynamic aspects of a given system. We start the specification of the structural design part by declaring the several sorts: `Sorts NFP_duration Method`. Then, we define all classes contained the design, as Real-time units or passive protected units regarding their roles in the design. Stereotyping, inheritance are also defined.

```
class Speed .
class Rotation_Sensor | sensor_description : String ,
                        periodicity : NFP_duration .
class Wheel | statue : String , description : String .
class Speed_Controller . class Speed_Notify .
class Speed_Client . class Cruise_Controller .
class Distance_Client . class Distance .
class Distance_Controller . class Distance_Notify .
class Laser | sensor_description : String , periodicity : NFP_duration  .
class Obstacle | statue : String , description : String .
```

RtUnit and PpUnit stereotypes are considered as inheritance in Maude :

```
class RtUnit | isDymamic : Bool , ismain : Bool ,
               poolsize : Float , main : String .
Subclass Rotation_Sensor Speed_Controller
Distance_Controller Laser < RtUnit .
class PpUnit | concpolicy : String . subclass Speed Distance < PpUnit .
```

The different methods in classes are specified as a Maude operations. Furthermore, we define a `Methods` operation allowing, for each method, to know the class to which it belongs.

```
ops Methods getValue Speed_Subscribe .
Speed_Unsubscribe Distance_Subscribe .
Distance_Unsubscribe getnext  : -> Set{Method} .
```

The specification of the different associations between classes (undirected association is considered as a two associations, one in each direction) is achieved in the following Maude code. Each association is specified as a Maude operation taking as parameter the first class and as result the second one. The multiplicity is also specified by `Set` and `NeSet`sorts for denoting respectively (1..*) and (0..*) multiplicities.

```
--- Associations definition as Maude operation
op Speed_Sensor  : Speed ->   Rotation_Sensor .
op Speed_Measure : Rotation_Sensor  -> Speed .
op Speed_C       : Speed_Controller -> Speed .
op Notified      : Speed_Controller -> Set{Speed_Notify} .
op Use_Speed     : Speed_Controller -> NeSet{Speed_Client} .
--- ...
eq  Methods(Speed) = getValue .
eq  Methods(Speed_Controller) =  Subscribe Unsubscibe .
eq  Methods(NotifieV) =  getnext .
```

A dynamic design represented as a sequence diagram and it shows the execution scenario of an activity. In our example, this diagram models the speed/distance capturing activity. The specification of this model is divided in two parts. In the first one, we define all sorts, operations and equations requested for each activity. However, the second is specific for each activity (capturing activity). The important element in this model is the signal. Thus, we define a sort `Signal` and all temporal features are defined as Maude operations upon it.

```
--- General Specification (classes instantiation and temporal features)
sorts Time Signal .
vars  O : Oid  C : class .
op Instance : class -> Oid .
op operation : Signal -> Method .
ops Trigger Targetc : Signal -> Oid .
ceq Target(S : Signal) = < O : C | > if operation(S) in Methods(C) .
op periodic : Signal -> Bool .
op periodVal : Signal -> Float .
ceq periodVal(S : Signal) = v : Float if periodic(S) .
op Time_ref : Signal -> Value .
```

```
op relDl : Signal -> Float .
op Rvalue : Time -> Float .
--- A capturing (Speed and Distance) activity.
Vars  Cruise_C Speed_C Distance_C Notify_Sp Notify_D
      Sp Dis  Rot_Sens Las : Oid .
eq Instance(Cruise_Controller) = < Cruise_C | isDymamic :
   false   ismain : false poolsize : 10 > .
eq Instance(Speed_Client) = < Speed_C | isDymamic : false
   ismain : true  poolsize : 10 main : Speed_Subscribe > .
eq Instance(Distance_Client) =  < Speed_C | isDymamic : false
   ismain : true poolsize : 10  main : Distance_Subscribe >.
eq Instance(Speed_Notify) :   Notify_Sp .
eq Instance(Distance_Notify) : Notify_D  .
eq Instance(Speed) =  < Sp | concpolicy : garded > .
eq Instance(Distance) =  < Dis | concpolicy : garded > .
eq Instance(Rotation_Sensor) = < Rot_Sens |
           isDymamic : false poolsize : 10 > .
eq Instance(Laser) = < Las | isDymamic   false poolsize : 10 > .
```

An algebraic semantic is associated to the Signal term `GETVALUE_S` through the following equations.

```
var t1 : Time  .
op GETVALUE_S : -> Signal .
eq operation (GETVALUE_S)  = getValue .
eq Trigger (GETVALUE_S)    = Speed_C .
eq Target (GETVALUE_S)     = Sp .
eq Time_ref (GETVALUE_S)   = t1 .
eq periodic (GETVALUE_S)   = TRUE .
eq periodVal(GETVALUE_S)   = 20 .
eq relDl (GETVALUE_S)      = 3.3 .
```

Some rewriting rules are added to Maude specification in order to manage temporal constraints. The following Maude declarations express the essential part.

```
msg Speed_Subscrib_Call : Oid Oid -> Msg [ctor] .
msg Distance_Subscrib_Call : Oid Oid -> Msg [ctor] .
vars C S Not_C : Oid .
op Speed_Subscrib_Signal : -> Signal .
eq operation(Speed_Subscrib_Signal) = Speed_Subscibe .
eq Trigger (Speed_Subscrib_Signal)  = Cruise_C .
eq Target (Speed_Subscrib_Signal)   = Speed_C .
Speed_Subscrib_Call (Cruise_C , Speed_C) .
rl[Speed_Sub] < C : Cuise_Controller > < S : Speed_Controller >
 Speed_Subscrib_Call => < C : Cuise_Controller >
 < S : Speed_Controller > < N : Speed_Notify > .
--- To ensure that a time constraint is verified
msg satisfy : Signal Signal Float -> Bool .
crl [ satisfy ] satisfy ( S1 : Signal , S2 : Signal , T : Flaot)
    if Rvalue( (Time_ref(S2) + reldl(S2)) - Time_ref(S1)) < T .
Satisfy(GETVALUE_S , GETNET_S , 5) .
```

5 Discussion and Conclusion

In the literature, we can find several work on meta-modelling approaches to define languages for design patterns. These works are in general based on UML and they aim to define a common model to all patterns in order to integrate them in CASE tools for assisting the designers (code generation or detection of patterns within a design for example).

Here, we can cite DPML (Design Pattern Modeling Language)[8] which defines a meta-model and a notation for specifying design pattern solutions and solution instances within object models. In the same context, Dae-Kyoo Kim et *al.* [7] present an UML-based pattern specification language called the role-based meta-modeling language (RBML), allowing to support the development of precise pattern specifications that can be used for the development of pattern tools.

In the context of the formal specification, we can cite two significant works namely, the BPSL (Balanced Pattern Specification Language) [11] and LePUS (LanguagE for Patterns Uniform Specification) [6], they aim to formalize the structural and behavior aspects of design patterns. BPSL uses a subset of first-order logic (FOL) to formalize structural aspect of patterns, while the behavioral aspect is formalized in TLA (Temporal Logic of Actions). LePUS is a fragment of the monadic high-level order logic using a limited vocabulary of entities and relations to describe a design pattern by HOL formulae accompanied by a graphic representation in order to facilitate its understanding.

In a previous work [4], we have proposed a rewriting logic based meta-model approach to formalize design pattern solutions and their instantiations. Our proposed meta-model includes all the common elements of design patterns, so any design pattern can be expressed in terms of this meta-model.

In this work, we are interest to formalize designs based on the real-time design patterns. Thus, we use first patterns instantiation and composition to generate a given design and repent it in UML-MARTE profile. This will permit us to consider the temporal properties and constraints of this RT pattern-based design. In the second time, we embbed in Maude language the representation result of the above design.

Our approach differs in two ways from the above cited works. Firstly, we deal with the real-time design patterns (especially those defined in [3]) and we consider also the temporal properties and constraints. Secondly, we use a common formalism (namely the RL logic) to specify both the structural and behavior aspects of design patterns. The encoding of models in Maude provides executable programs that can be subject to several analysis and verification.

This work is mainly a feasibility study for the proposed approach. We intend to extend the present work in two ways. The first one is to define a profile or a meta-model for real-time patterns to generate all possible patterns. Thus, this will serve to define a pattern instantiation mechanism to generate all possible solutions in conformity with their patterns. The second one is to formalize the defined meta-model and the instantiation mechanism, while ensuring formally the pattern-instance conformity. For this purpose, we plan to explore the RT-Maude (an extension of Maude for specifying and analyzing the real-time and

the hybrid systems) to encode the specification that will be more suitable to perform analysis and verification of the system proprieties.

References

1. Rekhis, S., Bouassida, N., Duvallet, C., Bouaziz, R., Sadeg, B.: A process to derive domain-specific patterns: Application to the real time domain. In: Catania, B., Ivanović, M., Thalheim, B. (eds.) ADBIS 2010. LNCS, vol. 6295, pp. 475–489. Springer, Heidelberg (2010)
2. Clavel, M., Durán, F., Eker, S., Lincoln, P., Martí-Oliet, N., Meseguer, J., Talcott, C. (eds.): All About Maude - A High-Performance Logical Framework. LNCS, vol. 4350. Springer, Heidelberg (2007)
3. Douglass, B.P.: Real-time design patterns: robust scalable architecture for real-time systems. The Addison-Wesley object technology series. Addison-Wesley, Boston (2003)
4. Douibi, H., Boukhelfa, K., Belala, F.: A rewriting logic-based meta-model for design patterns formalization. In: PATTERNS 2011: The Third International Conferences on Pervasive Patterns and Applications, pp. 84–89 (2011)
5. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley Longman Publishing Co., Inc., Boston (1995)
6. Gasparis, E.: Lepus: A formal language for modeling design patterns. In: Taibi, T. (ed.) Design Pattern Formalization Techniques, pp. 357–372. IGI Global (2007)
7. Kim, D.-k., France, R., Ghosh, S., Song, E.: A uml-based metamodeling language to specify design patterns. In: Patterns, Proc. Workshop Software Model Eng (WiSME) with Unified Modeling Language Conf. (2003)
8. Mapelsden, D., Hosking, J., Grundy, J.: Design pattern modelling and instantiation using dpml. In: CRPIT 2002: Proceedings of the Fortieth International Conference on Tools Pacific, pp. 3–11. Australian Computer Society, Inc., Darlinghurst (2002)
9. Meseguer, J.: Rewriting logic as a semantic framework for concurrency: a progress report. In: Sassone, V., Montanari, U. (eds.) CONCUR 1996. LNCS, vol. 1119, pp. 331–372. Springer, Heidelberg (1996)
10. Omgmarte.org. The uml profile for marte: Modeling and analysis of real-time and embedded systems (2015), http://www.omgwiki.org/marte, http://www.omgwiki.org (Last viewed January 2015)
11. Taibi, T., Ngo, D.C.L.: Formal specification of design patterns - a balanced approach. Journal of Object Technology 2(4), 127–140 (2003)

Author Index